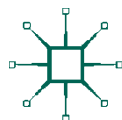




EDITED BY
KAREN P. CORRIGAN
ADAM MEARNES

CREATING & DIGITIZING LANGUAGE CORPORA

VOLUME 3:
DATABASES FOR
PUBLIC ENGAGEMENT



Creating and Digitizing Language Corpora

Karen P. Corrigan • Adam Mearns
Editors

Creating and Digitizing Language Corpora

Volume 3: Databases for Public Engagement

palgrave
macmillan

Editors

Karen P. Corrigan
School of English Literature, Language &
Linguistics, Newcastle University,
Newcastle Upon Tyne, United Kingdom

Adam Mearns
School of English Literature, Language &
Linguistics, Newcastle University,
Newcastle Upon Tyne, United Kingdom

ISBN 978-1-137-38644-1 ISBN 978-1-137-38645-8 (eBook)
DOI 10.1057/978-1-137-38645-8

Library of Congress Control Number: 2016952563

© The Editor(s) (if applicable) and The Author(s) 2016

The author(s) has/have asserted their right(s) to be identified as the author(s) of this work in accordance with the Copyright, Designs and Patents Act 1988.

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Cover illustration: © Carol and Hike Wemer / Alamy Stock Photo

Printed on acid-free paper

This Palgrave Macmillan imprint is published by Springer Nature
The registered company is Macmillan Publishers Ltd.

The registered company address is: The Campus, 4 Crinan Street, London, N1 9XW, United Kingdom



This book is dedicated to the memory of Lisa Lena Opas-Hänninen (1957–2013) who had been planning a chapter in the volume devoted to the LICHEN project. This would have been based on her presentation ‘Tools for making the most out of your multimodal data’ at the Corpora Galore workshop on 31 May 2011 at Newcastle University when undertaking an Erasmus teaching exchange visit there. LICHEN was funded by grants from the Emil Aaltonen Foundation and the Research Council of Norway and was part of a larger research network focusing on minority languages in the circumpolar region, coordinated by the University of Oulu where Lisa Lena was the head of English Philology. Research teams at Oulu and the University of Glasgow (where Lisa Lena held an honorary research fellowship) developed LICHEN in collaboration with another group of researchers, led by Bill Kretzschmar, working on the Digital Archive of Southern Speech (DASS) at the University of Georgia. It was an electronic framework, that is, a type of toolbox, for handling multimodal data and was tested using the Oulu Archive of Minority Languages in the North containing samples from the Kven, Meänkieli, Veps and Karelian languages, as well as DASS. The project’s key goals were to create a digital platform for collecting, managing and then

exploiting corpora with the primary objective of promoting the linguistic confidence and self-image of speakers from circumpolar communities. As Lisa Lena demonstrated in her Corpora Galore presentation, LICHEN's underlying premise is that culture and language are just as vital to populations as ecological or social concerns, and with this in mind the tools were developed with the aim of enhancing public engagement with the corpora. This theme runs throughout this entire volume and was an ideal that was very close to Lisa Lena's own views on the importance of improving access to the corpus data that the public contribute, so that everyone benefits—not just the academics.

Foreword: Doing the Right Thing

Academic conferences rarely produce epiphanies but, for me, the tenth *Methods in Dialectology* conference, held at Memorial University, St John's, Newfoundland, in the summer of 1999 did just that. I had gone there to present a joint paper by myself and Karen Corrigan, in which we reported some of the first research carried out on what was to become the NECTE corpus (Beal and Corrigan 1999). It was a wonderful conference,¹ but the epiphany came with the final plenary, delivered by Walt Wolfram. This plenary, entitled 'Principles of Donor Dialect Attribution', was a perfect fit with the conference theme of 'Historical Connections: Transported Varieties of English and their Origins', but the 'principle' that struck home for me was that of 'giving back' to the communities from which we take our data. As is apparent from the number of times it is cited in this volume, this principle, otherwise known as the 'principle of linguistic gratuity' (Wolfram 1993) has been fundamental to the understanding of impact and public engagement in sociolinguistics and other areas of linguistics. I had been out of sociolinguistics for a while, working with historical data, and this conference and my collaboration with Karen marked a return to the fold. Walt's plenary was inspirational, putting into words what I knew was just the right thing to do. As the chapter by Adam Mearns, Karen Corrigan and Isabelle Buchstaller

¹ See Ramisch (2000) for a brief report, and Clarke (1999) for a selection of papers.

in this volume demonstrates, this principle of doing right by the data we had inherited from other scholars, and by the community whose voices made up that data, was to inform all the projects carried out by the teams at Newcastle University working on the NECTE and DECTE corpora.

Of course, sociolinguistics and sociolinguists have always been ‘outward-facing’: Wolfram’s ‘principle of linguistic gratuity’ was preceded by Labov’s ‘principle of debt incurred’ (1982). In their introduction to an excellent collection of papers on impact in sociolinguistics, Lawson and Sayers (2016) ask:

Isn’t there a strong tradition of sociolinguistic work that has aimed, amongst other things, to fight language discrimination, to stand up for downtrodden minorities, and to champion linguistic equality? Doesn’t an emancipatory ethos resonate right from the earliest sociolinguistic research?

The questions are rhetorical, and the answer to both is *yes*: what could be more impactful than Labov’s early works such as *The Logic of Non-Standard English* (1969) and *Language in the Inner City* (1972)? Yet the demands of academic life are such that we constantly need to be reminded of these principles. On the one hand, there are the demands of institutional bean-counters obsessed with a different kind of ‘impact’: the impact factors of journals, measured in terms of citations in other journals, none of which are likely to have any impact beyond the academy.² On the other hand, the quest for scientific rigour in our discipline has produced more sophisticated methods of statistical analysis—quite proper and necessary, especially when dealing with the large data sets involved, but bringing with them the danger of forgetting the human beings behind the numbers.

Several of the contributors to this collection tell stories of achieving impact against the odds, combating what Kendall and Wolfram term ‘the problem of institutionalization’. When promotions, or even the allocation of time for research, depend on publishing in the *best* journals and time and effort spent on community engagement or data preservation are not recognized or supported, it can be hard for the overstretched researcher to prioritize these activities. Even getting a community-focused

²Unless, of course, they are made accessible to non-experts in the resumés provided by Cheshire and Fox in their Linguistics Research Digest!

project off the ground can prove challenging. Clarke refers to ‘the difficulty of finding sufficient funding for a project that focused primarily on the preservation of cultural heritage rather than on research as generally understood by the scientific community’ (p.109), and several of these projects have begun on a shoestring and/or depended on serendipitous changes in research council priorities. The projects described here have succeeded thanks to the initiative, inventiveness and entrepreneurship of the researchers involved. The fact that they have succeeded and the practical suggestions provided in these chapters should provide inspiration and help to readers planning their own projects.

The previous paragraph might give the impression that our contributors are all battle-scarred heroes, but the chapters in this volume also bear witness to the personal and professional rewards involved in doing right by our data and the communities from which it is taken. On a personal level, it is both refreshing and challenging to engage with people outside the narrow confines of the linguistics lab: we all need to *get out more!* Working with professionals in other disciplines, such as archivists, librarians and teachers, can give us new skills and perspectives, whilst engaging in genuine dialogue with non-linguists challenges our entrenched views. One important point made by several of the authors here is that ‘giving back’ is a two-way process. As Kendall and Wolfram put it, ‘the relationship between research and outreach can, and should be, bilateral and synergistic rather than one-dimensional—from research to outreach’ (p. 145). The very term ‘outreach’ has an evangelical flavour to it, as if we are missionaries sent to enlighten the natives rather than partners in a common endeavour. However, if we genuinely see our participants as equal partners, we need to respect and take seriously their views on language, which can be difficult when those views seem prescriptive or erroneous. Having the humility to concede that we might not always be right can lead to new insights.

Although historically, as many of our contributors testify, there have been difficulties in financing and sustaining community-focused projects, there are signs that the climate is becoming more favourable. In the UK, the introduction of ‘impact’ as a measurable factor in the Research Excellence Framework’s exercise of 2014 has increased institutional support for projects that engage with the public, and brought recognition to colleagues who were able to provide credible narratives for impact

case studies. Let us hope that this does not turn out to be a passing fad, and that the benefits of impactful research continue to be recognized and rewarded. On another level, as Kretzschmar's contribution suggests, the zeitgeist is becoming more sympathetic to projects that celebrate the distinctiveness of communities and their languages and dialects. UNESCO's definition of intangible heritage includes language and dialect, and the backlash against globalization has led to an appreciation of the local which provides opportunities for linguists to collaborate with tourism and heritage professionals. Such projects can actually make a difference to the economic viability of communities struggling with the loss of traditional industries: to return to my starting point, Walt Wolfram's contribution to the tourist industry of Okrakoke is a shining example of this, but Kretzschmar points to intriguing possibilities for the future.

Since that epiphanic encounter in St John's, I have contributed to and promoted a range of activities whose institutional labels have changed every few years, from *innovation* to *knowledge transfer* to *knowledge exchange* to *public engagement* and *impact*. At times, it has been hard to convince colleagues that such activities are relevant to the humanities and social sciences and that they are not a distraction from the serious business of research. I must confess to being tempted to say *I told you so* when impact was included in the 2014 REF! The contributions to this volume are a vindication of those who stuck to the 'principle of linguistic gratuity' in the face of such opposition, but they are also a refutation of the naysayers who consider community-based projects as *not serious research*: as Clarke puts it, 'while our aims are to promote cultural heritage by engaging current generations, we have not done this at the expense of academic rigour' (p. 102).

The publication of this volume is, in itself, a sign that attitudes to public engagement in research are changing, and credit must be given to Palgrave Macmillan for having the vision to work with the editors and promote this collection. I am sure that it will become essential reading for researchers at all stages, providing as it does inspiration and practical advice for anyone embarking on a corpus-based project. Most importantly, it will help to increase the effectiveness of community-based projects, bringing benefits to linguists and non-linguists alike. Read it, and then get out there!

References

- Beal, Joan C. and Karen P. Corrigan. 1999. Comparing the present with the past to predict the future for Tyneside English. Paper presented at *Methods in Dialectology X*, St. John's, Newfoundland, Memorial University of Newfoundland, August 1999.
- Clarke, Sandra (ed.). 1999. *Xth International Conference on Methods in Dialectology: Methods X, Memorial University, St. John's, Nfld., Canada, August 1999*. St John's: Memorial University of Newfoundland.
- Lawson, Robert and Dave Sayers. 2016. Introduction: A history of impact. In *Sociolinguistic Research: Application and Impact*, Robert Lawson and Dave Sayers (eds). London: Routledge.
- Labov, William 1969. The logic of nonstandard English. In *Georgetown Monograph Series on Language and Linguistics No. 22*, James E. Alatis (ed.), 1–43. Washington, DC: Georgetown University Press.
- Labov, William. 1972. *Language in the Inner City: Studies in the Black Vernacular*. Philadelphia, PA: University of Pennsylvania Press.
- Labov, William. 1982. Objectivity and commitment in linguistic science. *Language in Society* 11: 165–201.
- Ramisch, Heinrich. 2000. Conference Report: 10th International Conference on *Methods in Dialectology* (Methods X), Memorial University of Newfoundland, St John's Nfld, Canada, August 2–6 1999. *Dialectologia et Geolinguistica* 8, 124-126.
- Wolfram, Walt. 1993. Ethical considerations in language awareness programs. *Issues in Applied Linguistics* 4: 225-255.
- Wolfram, Walt. 1999. Principles of Donor Dialect Attribution. Plenary lecture presented at the 10th International Conference on *Methods in Dialectology*, Memorial University of Newfoundland, St John's, Nfld, Canada, August 2–6 1999.

Joan C. Beal
University of Sheffield,
Sheffield, United Kingdom

Contents

1 Taming Digital Texts, Voices and Images for the Wild: Models and Methods for Handling Unconventional Corpora to Engage the Public	1
<i>Karen P. Corrigan and Adam Mearns</i>	
Part I Corpora for Education and Heritage	23
2 Migration Databases as Impact Tools in the Education and Heritage Sectors	25
<i>Carolina P. Amador-Moreno, Karen P. Corrigan, Kevin McCafferty, and Emma Moreton</i>	
3 Engaging Users of Scottish Online Language Resources	69
<i>Wendy Anderson and Carole Hough</i>	
4 From Legacy Regional Language Materials to Public Engagement: The Interactive Online Dialect Atlas of Newfoundland and Labrador	99
<i>Sandra Clarke</i>	

5	Engagement Through Data Management and Preservation: The North Carolina Language and Life Project and the Sociolinguistic Archive and Analysis Project	133
	<i>Tyler Kendall and Walt Wolfram</i>	
6	Roswell Voices: Community Language in a Living Laboratory	159
	<i>William A. Kretzschmar, Jr</i>	
7	The Diachronic Electronic Corpus of Tyneside English and The Talk of the Toon: Issues in Preservation and Public Engagement	177
	<i>Adam Mearns, Karen P. Corrigan, and Isabelle Buchstaller</i>	
8	Language Learning at Your Fingertips: Deploying Corpora in Mobile Teaching Apps	211
	<i>Seth Mehl, Sean Wallis, and Bas Aarts</i>	
Part II	Corpora for Continuing Professional Development	241
9	Locating People with Their Language: An Applied Linguistics Course Using Linguistic Microvariation Databases and Tools	243
	<i>Sjef Barbiers</i>	
10	From Sociolinguistic Research to English Language Teaching	265
	<i>Jenny Cheshire and Susan Fox</i>	

11	Analysing Spoken Discourse in University Small Group Teaching	291
	<i>Steve Walsh and Dawn Knight</i>	
12	The Wellington Language in the Workplace Project: Engaging with the Research and Wider Communities	321
	<i>Bernadette Vine</i>	
	Index	347

Author Bios

Bas Aarts is Professor of English Linguistics and Director of the Survey of English Usage at UCL. His research interest is in the field of syntax, more specifically verbal syntax. His recent publications include: *Syntactic Gradience* (2007), *Oxford Modern English Grammar* (2011), *The English Verb Phrase* (2013, edited with J. Close, G. Leech and S. Wallis), *Oxford Dictionary of English Grammar* (2nd edition, 2014; edited with S. Chalker and E. Weiner), as well as articles in books and journals. He is a founding editor of the journal *English Language and Linguistics*. University College London, London, UK.

Carolina P. Amador-Moreno graduated from the University of Ulster and the University of Extremadura, where she did a European doctorate in 2002. She is currently Senior Lecturer in English at the University of Extremadura. She has held different teaching positions at the University of Limerick (Department of Languages and Cultural Studies), and was also lecturer in Hiberno-English at University College Dublin (English Department). Current research projects include CONVAR (*Contact, Variation and Change*), in collaboration with Kevin McCafferty, at the University of Bergen. University of Extremadura, Badajoz, Spain.

Wendy Anderson is Senior Lecturer in English Language at the University of Glasgow. Her research and teaching interests include semantics, metaphor, translation and intercultural language education.

She is the editor of *Language in Scotland: Corpus-based Studies* (2013), co-editor with Ellen Bramwell and Carole Hough of *Mapping English Metaphor through Time* (2016) and co-author with John Corbett of *Exploring English with Online Corpora* (2009, Palgrave). She is the associate editor of *Journal of Digital Scholarship in the Humanities*. University of Glasgow, Glasgow, UK.

Sjef Barbiers is a senior researcher at the Meertens Instituut in Amsterdam and Professor of Variation Linguistics at Utrecht University. His field of specialization is syntactic microvariation, in particular in Dutch. He was Principal Investigator of the Syntactic Atlas of the Dutch Dialects project (SAND) and the European Dialect Syntax project (Edisyn). Meertens Instituut, Amsterdam, Netherlands Utrecht University, Utrecht, Netherlands.

Isabelle Buchstaller is Professor for Varieties of English at Leipzig University. Her main research interest is language variation and change. She has worked extensively on the DECTE corpus, considering in particular longitudinal morphosyntactic changes, such as intensifiers, quotation, stative possession and future time reference. Her recent research examines the extent to which the grammars of individual speakers can change in relation to ongoing community-wide changes. Leipzig University, Leipzig, Germany.

Jenny Cheshire is Professor of Linguistics at Queen Mary, University of London. She is a Fellow of the British Academy, and editor of the journal *Language in Society*. Her research is on language variation and change, with a recent focus on Multicultural London English and Multicultural Paris French. Publications include *Variation in an English Dialect* (1982), *English around the World* (1991), *The Sociolinguistics Reader*, with Peter Trudgill (1998) and *Social Dialectology*, with David Britain (2003). Queen Mary University of London, London, UK.

Sandra Clarke is Professor Emerita of Linguistics at Memorial University of Newfoundland. Her research deals with social and sociohistorical variation, with particular focus on Newfoundland and Canadian English, as well as the indigenous Algonquian varieties of Labrador. Her recent work has involved the mobilization of linguistic knowledge on regional varia-

tion in Newfoundland and Labrador English for public as well as academic audiences. Memorial University of Newfoundland, St. John's, Canada.

Karen P. Corrigan has lectured at University College, Dublin, and the Universities of Edinburgh and York. She is currently Professor of Linguistics and English Language at Newcastle University. She co-edited the two previous volumes in this series and has also recently published *Irish English, Volume 1: Northern Ireland* (2010). Newcastle University, Newcastle upon Tyne, UK.

Susan Fox is a lecturer at the University of Bern, Switzerland. She is a sociolinguist whose research interests are language variation and change—especially in urban multicultural contexts—multiethnolects, language and dialect contact, the impact of immigration on language change and the language of adolescents from a variationist perspective. Her research has mainly focused on the social and historical contexts that have led to the variety of English that is spoken today in London. University of Bern, Bern, Switzerland.

Carole Hough is Professor of Onomastics at the University of Glasgow. Her research interests include name studies, semantics, historical linguistics and lexicography. She is editor of *The Oxford Handbook of Names and Naming* (2016) and co-editor with Wendy Anderson and Ellen Bramwell of *Mapping English Metaphor Through Time* (2016). She is a former president of the International Council of Onomastic Sciences and the International Society of Anglo-Saxonists. University of Glasgow, Glasgow, UK.

Tyler Kendall is Associate Professor of Linguistics at the University of Oregon, USA. Much of his work focuses on corpora and computational approaches to the study of language variation and change. He is the developer of several sociolinguistic software projects, including the Sociolinguistic Archive and Analysis Project (SLAAP) and the *Vowels.R* package for the R programming language. He is author of the book *Speech Rate, Pause, and Sociolinguistic Variation: Studies in Corpus Sociophonetics* (2013, Palgrave Macmillan). University of Oregon, Eugene, OR, USA.

Dawn Knight is Senior Lecturer in Applied Linguistics at Cardiff University. Her research interests lie in the areas of corpus linguistics, discourse analysis, lexico-grammar, digital interaction and non-verbal communication. The main contribution of Knight's work has been to pioneer the development of a new research area in applied linguistics: multimodal corpus-based discourse analysis. This has included the introduction of a novel methodological approach to the analysis of the relationships between language and gesture-in-use using corpora. Cardiff University, Cardiff, UK.

William A. Kretzschmar Jr. is the Willson Professor in Humanities at the University of Georgia. He is the editor of the American Linguistic Atlas Project, and his books *The Linguistics of Speech* (2009) and *Language and Complex Systems* (2015) lay the foundation for analysis of language as a complex system. University of Georgia, Athens, GA, USA, University of Glasgow, Glasgow, UK, University of Oulu, Oulu, Finland.

Kevin McCafferty is Professor of English Linguistics at the University of Bergen. His research interests are in the field of language variation and change, with a focus on Irish English. His most recent publications include *Pragmatic Markers in Irish English* (2015), co-edited with Carolina P. Amador-Moreno and Elaine Vaughan, and articles in *English Language and Linguistics* and *American Speech*. University of Bergen, Bergen, Norway.

Adam Mearns has taught at the Universities of Sheffield and Leeds and at Northumbria University. He is currently Lecturer in the History of the English Language at Newcastle University. Recent publications have focused on the dialect of Tyneside and the concept of the supernatural in Old English. Newcastle University, Newcastle upon Tyne, UK.

Seth Mehl is a research assistant at the University of Sheffield, where he investigates semantic and conceptual change in Early Modern English. He was previously a research assistant at University College London, where he worked on multiple knowledge transfer and pedagogical innovation projects, as well as in widening participation programmes. He completed his PhD in English at University College London. University of Sheffield, Sheffield, UK.

Emma Moreton is a senior lecturer at Coventry University, where she teaches modules in corpus stylistics and discourse analysis. She has a BA in literature from De Montfort University, Leicester, and an MPhil and PhD, both in corpus linguistics, from the University of Birmingham. Emma was a member of the academic team on the BT Digital Archives project (<http://www.digitalarchives.bt.com/web/arena>), and she was co-investigator on an AHRC funded project, 'Digitising experiences of migration' (www.lettersofmigration.blogspot.com). Coventry University, Coventry, UK.

Bernadette Vine is a research fellow on the Language in the Workplace Project (www.victoria.ac.nz/lwp) and corpus manager for the Archive of New Zealand English. Her research interests include workplace communication, leadership and New Zealand English. She is the author of *Getting Things Done at Work: The Discourse of Power in Workplace Interaction* (2004) and is co-author, with Janet Holmes and Meredith Marra, of *Leadership, Discourse and Ethnicity* (2012). Victoria University of Wellington, Wellington, New Zealand.

Sean Wallis is a senior research fellow in Corpus Linguistics in the Survey of English Usage at UCL and has published on computing, artificial intelligence, corpus linguistics and statistics. The developer of the ICECUP software, he oversaw the completion of the ICE-GB and DCPSE corpora. Recent publications include contributions in the *English Verb Phrase* (2013), which he co-edited. He also writes the 'corp.ling.stats' statistics blog (<http://corplingstats.wordpress.com>). UCL, London, UK.

Steve Walsh is Professor and Head of Applied Linguistics in the School of Education, Communication and Language Sciences, Newcastle University. He has been involved in English Language Teaching for more than 30 years in a range of overseas contexts. His research interests include classroom discourse, teacher development and second-language teacher education. Newcastle University, Newcastle upon Tyne, UK.

Walt Wolfram is William C. Friday Distinguished University Professor at North Carolina State University, where he also directs the North Carolina Language and Life Project. He has pioneered research on social

and ethnic dialects since the 1960s and authored or co-authored more than 20 books and 300 articles. One of his enduring concerns is the application of sociolinguistic information for the public, ranging from students and the university community to the general public. North Carolina State University, Raleigh, NC, USA.

List of Figures

Fig. 2.1	An advertisement for the Columbia et al. ships, Belfast to USA and Canada (<i>top</i>), and a bill for protection and relief of destitute poor evicted from dwellings in Ireland (<i>bottom</i>)	28
Fig. 2.2	Digital copies of: (a) PRONI T3028/B/5(2)—a letter written by Lewis Reford in Newburgh, Orange County, York State, to Fanny Reford, County Antrim on July 15th, 1849 (<i>left</i>); and (b) PRONI D1819/3—a letter written by Mary Quin, Barrytown, New York to her sisters in Stewartstown, Co. Tyrone on January 1st, 1873	41
Fig. 2.3	Document analysis (letter by Julia Lough, Winsted, 1891 to her mother)	44
Fig. 2.4	Example mark-up (person, location and date)	47
Fig. 2.5	(a) An example of the raw metadata from the letter collection held at the IHRC, University of Minnesota; (b) TEI compliant metadata	50
Fig. 2.6	Visualization 1 (<i>top</i>): The location and movement of migrants. Visualization 2 (<i>bottom</i>): Letter writing networks	51
Fig. 2.7	Sample of foamex board from the Herbert exhibition	54
Fig. 2.8	Extracts from the <i>From Home to Here</i> booklet	59
Fig. 2.9	Extract from Séan Ó Dúbhda's response to the 1955 Irish Folklore Commission's Questionnaire	60
Fig. 2.10	Extract 3 from the <i>From Home to Here</i> booklet	61

xxiv List of Figures

Fig. 3.1	Screenshot of SCOTS corpus, showing a search for the word-form <i>dour</i> , with associated Google Maps results	74
Fig. 3.2	Screenshot from CMSW website of real time blog of Thomas Crawford's journey from Scotland to Australia	78
Fig. 3.3	Screenshot of <i>Scots Words and Place-names</i> website, showing sample item <i>hair</i> in the place-name glossary	79
Fig. 3.4	Screenshot of beta-version of Metaphor Map online resource, showing the categories with metaphorical links to <i>Life</i>	83
Fig. 4.1	<i>Left:</i> A map illustrating the communities used in the pronunciation and grammar section of the <i>Dialect Atlas of Newfoundland and Labrador</i> . <i>Right:</i> A map illustrating the communities used in the vocabulary section of the <i>Dialect Atlas of Newfoundland and Labrador</i>	104
Fig. 4.2	Secondary worksheet (partial) for postvocalic /l/, Paddock survey	106
Fig. 4.3	Regional distribution of the term <i>barm</i> ('old-fashioned yeast') in Newfoundland and Labrador	114
Fig. 4.4	Response breakdown for the term <i>piper</i> ('tin kettle') in the community of Port de Grave	115
Fig. 4.5	TIE rounding, with illustrative tokens from Round Harbour. Inset: Round Harbour sound clip and transcription for the token <i>nine</i>	118
Fig. 4.6	Information box for the 'Long A vowel' variable	121
Fig. 4.7	Interactive games from the online Atlas	122
Fig. 4.8	Partial view of a user contribution form	123
Fig. 5.1	Four presentations available in SLAAP of the same transcript data (from Kendall 2007)	138
Fig. 5.2	<i>Praat</i> TextGrid for the transcript shown in Fig. 5.1	140
Fig. 5.3	SLAAP screenshot showing a transcript line with phonetic data	141
Fig. 5.4	Example of QR use with printed text	149
Fig. 5.5	Website and Chapter Media for <i>Talkin' Tar Heel</i>	151
Fig. 6.1	<i>Roswell Voices</i> on the Roswell Folk and Heritage Bureau website	164
Fig. 6.2	<i>Roswell Voices</i> on the ENoLL website	171
Fig. 7.1	DECTE: components and phases of development	182

Fig. 7.2	Extracts from the NECTE2 Orthographic Transcription Protocol (<i>top</i>) and a completed NECTE2 Transcription File (<i>bottom</i>)	188
Fig. 7.3	The home page of <i>The Talk of the Toon</i>	196
Fig. 7.4	The interview interface page for decten2y10i009	197
Fig. 7.5	The interview interface page for decten2y10i007, showing the anonymization of personal names	199
Fig. 7.6	The interview interface page for decten2y10i009, with the Sport theme selected	202
Fig. 7.7	The main Themes page, showing the results for Entertainment & Culture (<i>left</i>) and the Picture results for the Entertainment & Culture theme (<i>right</i>)	203
Fig. 7.8	A question from one of the <i>Talk of the Toon</i> interactive quizzes (<i>left</i>) and materials from the Discovery Museum workshops (<i>right</i>)	205
Fig. 7.9	<i>The Talk of the Toon</i> booklet	206
Fig. 8.1	The form factor challenge: iGE on a tiny Sony Xperia Mini Android phone (<i>left</i>) and Motorola Zoom tablet (<i>right</i>)	217
Fig. 8.2	Sketch of app structure: navigation tools (<i>left</i>) and course content (<i>right</i>). The course consists of a series of ‘chapters’ containing sequential explanation and exercise pages	219
Fig. 8.3	A ‘spot the noun’ exercise in iGE (iPhone). The exercises can be performed in horizontal or vertical orientation	222
Fig. 8.4	Left: an example AWE content page (iPhone). Right: an exercise in AWE on the iPhone to select the most appropriate transition. The option <i>But also ...</i> has just been selected	224
Fig. 8.5	ESP’s Spelling Practice module. The Progress Report (<i>right</i>) shows the user has completed 23 of the first 25 words, with a best score in that round of 100 per cent. The panel below lists the words learned in order of difficulty	230
Fig. 9.1	The MIMORE search tool	256
Fig. 9.2	MIMORE virtual collection	257
Fig. 10.1	The English language teaching resources archive	270
Fig. 10.2	The Linguistics Research Digest: first part of a summary	279
Fig. 10.3	First page of one of the Databank documents	280
Fig. 10.4	Example of a Language Investigation	282
Fig. 11.1	Extract from NC0030 Business game (0.54–1.27)	304
Fig. 11.2	Extract from NC0030 Business game	307

List of Tables

Table 2.1	Person, location, date	46
Table 4.1	Extract from the ‘House and household’ section of the lexical questionnaire	113
Table 4.2	Examples of wording changes in pronunciation and grammatical queries	120
Table 6.1	Phase I and Phase II interviews	167
Table 7.1	The current composition of DECTE	193
Table 10.1	Folders in the Databank of spoken English	271
Table 11.1	Contents of the NUCASE subcorpus	297
Table 11.2	The top 20 words in each of the subjects	299
Table 11.3	The most frequent pronouns and their (relative) frequencies across each of the subjects	300
Table 11.4	The most frequently used ‘discourse bin’ terms featured in the Business, Bioinformatics and Marine Engineering data	301

1

Taming Digital Texts, Voices and Images for the Wild: Models and Methods for Handling Unconventional Corpora to Engage the Public

Karen P. Corrigan and Adam Mearns

1 Stimulus for the Volume and Its Overarching Aim

This volume is the third in a series of books published by Palgrave Macmillan which focus on establishing guidelines for the creation and digitization of language corpora that are unconventional in some respect (see Beal et al. 2007a, b). Volume 3 is dedicated to the issue of public engagement and questions of how linguists can and should make their corpora accessible for a broader range of uses and to a wider audience. Although in this regard the road to building a corpus is often paved with

The term ‘unconventional’ here relates to the distinction first articulated in Beal et al. (2007a, b) between large-scale standardized or conventional corpora like the *International Corpus of English* or COBUILD and smaller more specialized databases. These are often not devised at the outset as corpora strictly speaking since they initially arise from sociolinguistically oriented projects, but such resources can indeed be used as such providing they are ‘tamed’ in particular ways (Beal et al. 2007a: 1). See also D’Arcy (2011: 54–6) and Kendall (2011: 362–3).

K.P. Corrigan (✉) • A. Mearns
Newcastle University, Newcastle upon Tyne, UK

© The Editor(s) (if applicable) and The Author(s) 2016
K.P. Corrigan, A. Mearns (eds.), *Creating and Digitizing
Language Corpora*, DOI 10.1057/978-1-137-38645-8_1

good intentions, as Rickford (1993: 130) observes, these are frequently overtaken by ‘the less escapable commitments’ of teaching and further research. While this may be understandable, it is ‘not a picture, when we step back and view it, with which we can be proud’, since it means that ‘[m]ost of us fall short of paying our debts to the communities whose data have helped to build and advance our careers’ (Rickford 1993: 130). The importance of taking public engagement initiatives more seriously has generated considerable recent scholarly debate (especially amongst researchers in the arts, humanities and social sciences) as the so-called ‘impact agenda’ has taken hold particularly, though not exclusively, in UK higher education institutions (Martin 2011; Samuel and Derrick 2015; Lawson and Sayers 2016).¹ A key objective of this volume is to examine the evidence for the view that despite the new requirements by funding bodies (and ultimately governments) that corpora should have a dual purpose as data that is deployable for engagement as well as research, twenty-first-century corpus linguists who do just that are not following conventional practices within their discipline. A second goal is to demonstrate how the issues that purportedly stand in the way of developing what one might term ‘impactful corpora’ can be circumvented (as our contributors have done) with a little ingenuity and motivation. Another objective is to sketch what we consider to be best practices in creating corpora for public engagement by offering guidance on optimal methods by which such data (audio, text and still/moving images) can be created, digitized and subsequently exploited for public engagement projects.

To these ends, all the digital, community-oriented initiatives described in this volume are predicated on the premise that linguists should not engage solely in research that ‘produces or intensifies an unequal relationship between investigator and informants’ (Cameron et al. 1997: 145) but should instead be governed by the principles of ‘linguistic grauity’ (Wolfram 1993, 2012, 2013, 2016; Reaser and Adger 2007: 168; Wolfram et al. 2008); and ‘debt incurred’ (Labov 1982). Some of the chapters in the volume derive in part from presentations at a peer-reviewed

¹The introduction by Lawson and Sayers (2016) to their Routledge volume, which explores the possibilities for combining sociolinguistic research with the impact agenda, offers an excellent historical overview of how this ideology developed and its implications for scholarship from the 1980s to the present day.

workshop entitled *Dialect and Heritage Language Corpora for the Google Generation* which was organized by the editors for the 2011 *Methods in Dialectology XIV* Conference at the University of Western Ontario. Other papers were delivered at the *Corpora Galore: Applications of Digital Corpora in Higher Education Contexts* workshop also organized by Corrigan and Mearns at Newcastle University in May 2011. These papers are supplemented by invited contributions from key scholars who have an international reputation as creators of divergent digital materials relevant to the preservation, analysis and public dissemination of dialect and heritage language corpora.

1.1 How to Tame Digital Texts, Voices and Images for the Wild

A reviewer for our proposal to Palgrave Macmillan outlining the case for editing a volume on corpus creation that would focus on engagement rightly contended that ‘despite the requirement imposed by funding agencies that corpora should be constructed with public engagement in mind, this proves to be the exception rather than the rule’. There are three principal reasons why we consider this view to be justified: the diversity of aims between one corpus creation project and another; the very understandable desire amongst those who have given their blood, sweat and tears to collect the data and build a corpus from it to keep the resource for private use;² and the extent to which a corpus can ever be effectively anonymized for public access.

Corpora are created and digitized by academics for a wide range of purposes but their primary function is to address the particular hypotheses which underpin their research projects, whatever these may be. Thus, the basic processes of designing a corpus for research into the acquisition and structure of L2 phonetic and phonological systems will be rather different from those one might avail oneself of when building a corpus that will investigate divergence in the discourse marking systems across

² Often in the sense that there is no intention to share it with other scholars, let alone members of the public. On the distinction between public and private corpora, see also Bauer (2004) and D’Arcy (2011: 51–6).

varieties of French. Although there has been an appetite in several recent publications for normalizing approaches to corpus collection, data management and annotation strategies and for advocating more collaboration amongst researchers in this field (Kretzschmar et al. 2006; Beal et al. 2007a, b; Kendall 2011; Durand et al. 2014), it is difficult to unify an activity which has to satisfy very divergent and often conflicting research goals. Moreover, with very few exceptions (such as Amador et al. and Vine, this volume, as well as some of the projects described in Anderson and Hough) corpora are built first and foremost for mining by academics, and considering how they might be adapted for societal impact uses is generally an afterthought, that is *ex post* rather than *ex ante* in the terms of Samuel and Derrick (2015). With this fact in mind, some resourcefulness is needed on the part of scholars to repurpose their corpora for wider audiences. These can sometimes involve straightforward steps such as the conversion of the *Diachronic Electronic Corpus of Tyneside English* (DECTE) XML files into plain text so as to make them more accessible, as described in Mearns et al. (this volume). For other data sets, in order to convert one's corpus to a form that can engage a particular demographic of public end-users, the procedures are considerably more complex (see Norris 2001; Rowlands et al. 2008; Choudhrie et al. 2010). This would be the case, for example, when creating apps for smartphones and tablets that are built on the *Survey of English Usage* (SEU) corpus, which Mehl et al. (this volume) illustrate.

Wherever one stands politically on the impact agenda, there is no denying that it presents scholars with opportunities to devote research time to rethinking new end-uses like these for their data sets that may in the past have been considered nothing more than a vanity project on account of the media and wider public attention that doing so might have attracted. It is interesting to note in this regard that the new agenda can be considered a strong motivator for corpus design that can engage the public in this way. This is evidenced by the fact that the chapters in this volume by Anderson and Hough, Cheshire and Fox, Mearns et al. and Mehl et al. are all connected with Impact Case Studies submitted for the 2014 Research Excellence Framework exercise in the UK. In fact, only two of our contributions from UK authors are not linked with Impact Case Study submissions (Amador et al. and Walsh and Knight). In each case

this is likely to be due simply to the recent nature of their corpus-building activities which have not yet had the kind of gestation period required for the gathering of evidence to support societal impact claims of the type expected.³ The UK, of course, is not alone in promulgating the view that one's research cannot simply take an 'art for art's sake' stance. Hence Wolfram (2016: 87) notes that: 'Every research proposal submitted to the National Science Foundation in the United States requires a narrative section titled "Statement of Broader Impacts". Under this heading, the principal investigator is obliged to address the project's "benefits to society" and to "human welfare"'. The contribution to this volume by Barbiers examines how dialect atlas projects in the Dutch-speaking parts of the Netherlands and Belgium can be repurposed to deliver similar goals. As such, scholars globally are being encouraged to plan for and integrate outreach activities within their core research agenda, so a volume like this which demonstrates how such obligations can be met even with corpora that were not originally created for these purposes is not only timely but we hope will serve as a 'go-to guide' for obtaining awards to fund future impactful corpus-building initiatives.

This orientation of course may conflict with other goals that corpus linguists have, particularly those who have invested considerable time (and money) in creating the kind of smaller-scale unconventional corpora which are the focus of Beal et al. (2007a, b) as well as contributions to this volume. There are many good reasons to keep such initiatives private including 'scoop' avoidance, as Childs et al. (2011) put it, which is probably why this is their 'default' status (D'Arcy 2011: 55). Another issue, of course, is that granting wider access to one's hard-won corpus might also lead to original analyses being refuted or otherwise invalidated. These issues aside, there does, however, seem to be a welcome change afoot amongst corpus creators towards widening access—if not to the public then at least to other scholars (as the data sets deposited with the *Sociolinguistic Archive and Analysis Project* (SLAAP) described in Kendall and Wolfram here testify).

³This can be gathered from the published guidelines: 'an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia' (Research Excellence Framework 2011: 26).

This new mindset is partially fuelled by funding bodies who, naturally, want to add as much value as possible to research they have supported and also because they have readily adopted new policies embracing Open Access (Kendall 2008, 2011; Childs et al. 2011).⁴ Their demands can be met by having principal investigators lodge their corpus data at project end in suitable national repositories (such as Qualidata and the Oxford Text Archive (OTA) in the UK or the Linguistic Data Consortium (LDC) in the US) and indeed research councils in the UK have an expectation that they will do just that. We strongly endorse this kind of good practice amongst corpus linguists not least because we have personal experience of how one of the data sets which eventually formed a cornerstone of the *Newcastle Electronic Corpus of Tyneside English* (NECTE) initiative, namely the *Tyneside Linguistic Survey*, was only salvageable on account of the fact that one of the original project team had lodged materials from the project with the OTA in the 1970s (see Allen et al. 2007).

What of the cases, though, where private corpora were never intended to have public uses—an issue which is particularly problematic when the corpus in question derives from legacy materials as NECTE does? As Kretzschmar et al. (2006: 191) advocate there must nevertheless be ‘a clear pathway toward publication of our materials that preserves the rights of both our research participants and of the corpus builders’. That paper demonstrates how this has been done with ‘major success’ for the *Linguistic Atlas Project* (LAP). It will thus be useful in this context to rehearse the rights management practices of the LAP project since they offer an insight into a mechanism that could easily be adapted by others. The team, for example, developed a consent form that did not guarantee anonymity in all contexts to participants. They devised statements for initialling by informants reflecting the exact intentions of the project team with respect to different options for making the corpus data public. These ranged from online publication (where participants’ names or other personally identifying information are not released) to local public exhibitions (where speakers might indeed be named or identified). This is possible of course for contemporary corpus data like LAP, but what practices would we advocate in cases where the consent forms—if

⁴ See also: <http://www.rcuk.ac.uk/research/openaccess>.

there even were any in the first place—were completed some time ago and did not envisage wider access (such as publication on the web, for example)? Allen et al. (2007) describe a work-around in the form of accessing the corpus via a password-protected system, which was developed for NECTE and which relates exclusively to the corpus as it was originally conceived to be, that is primarily an academic resource. Mearns et al. (this volume) describe the anonymization protocol that we have had to adapt during subsequent phases of this corpus-building initiative that were instead public-facing in their orientation. Similar good practice is described in the contribution by Cheshire and Fox here which focuses, for instance, on the differential treatment which they advocate for different types of corpus file. Their spoken data is restricted to the project team and a very small number of bona fide researchers with whom they are closely associated. The text files, by contrast, have not only been shared with a wide range of academic colleagues but also form the backbone of their Language Archive designed particularly for public engagement purposes. However, as Cheshire and Fox note, the issues are complex and often require unique, community-specific solutions. The manner in which human subjects and copyright matters are problematized in this contribution and in others in this volume, as well as in previous research by Allen et al. (2007), Childs et al. (2011), Kendall (2011) and Kretzschmar et al. (2006), offers salutary lessons for us all.

An important take-home message from a majority of the corpus-building for public engagement initiatives described here is the time lag between data creation and their eventual application in real-world contexts. This necessitates developing sound protocols for the capture and preservation of metadata (machine-understandable information for the web) associated with the original resources and subsequent instantiations of them. While Clarke's dialect atlas, DECTE, the *North Carolina Language and Life* (NCLL) and the SEU corpora are probably the most extreme of our examples in this regard, even those projects like Cheshire and Fox's Language Archive already noted did not launch until several years after the first corpus-building phase of their London projects. As such, we recommend that all corpus creators need to reflect on the longer-term potential of exploiting their data for public outreach from the earliest stages of their projects—irrespective of any research goals.

Doing so will, for example, allow all the correct permissions to be in place for the widest possible range of engagement types and it will also encourage a corpus design strategy that is as flexible as possible so that the data can be adapted in time to come.

‘Sustainability’ is an important ‘buzzword’ with respect to corpus design, and indeed clarifying the planned steps for future-proofing is an explicit criterion which research councils apply when reviewing grant applications for corpus-building initiatives. Given the length of time already noted as a norm between the release of a corpus and its application for public engagement purposes, taking measures that will ensure the longevity of your data set should thus be taken more seriously than simply paying lip service to research council edicts. Since the vast majority of corpora described in this volume are either web-based in their entirety or use online technologies such as quick response (QR) coding we would recommend that impactful corpus builders ensure that they comply with the latest guidelines for web standards developed by the World Wide Web Consortium.⁵ Doing so will guarantee that their resources have the maximal shelf life possible. Ideally, of course, this will involve a programme of revision and updating long after the corpus has launched (and the grant has run out). Individuals may be in a position to do this while they remain employed at higher education institutions, and the contribution by Anderson and Hough outlines exactly how this is achieved for their resources at Glasgow. Other projects described in this volume have practices whereby their public outputs—in the form of apps, books, pamphlets and the like—are marketed and the profits are ploughed back into the upkeep of the corpus. These practices too, though, are dependent on individuals sustaining them over a period of time, which is less than ideal. In addition, therefore, to advocating the deposit of data sets with organizations such as the LDC or OTA, we would also recommend developing good relations with ‘Living Laboratory’ initiatives (Kretzschmar, this volume), university libraries and Information Technology groups which are best placed to steward major archiving projects like these in the longer term (see Day 2001; Smith et al. 2004; Perks 2011; Robertson 2011; Kendall and Wolfram, this volume).

⁵ See: <http://www.w3.org>.

As well as building corpora that can be adaptable from the perspective of the ongoing development of new technologies and software, we also recommend that they are designed so that they can be presented and repurposed for multiple modes of engagement uses. This will entail, for example, designs employing techniques such as the software created for the SLAAP initiative described in Kendall and Wolfram (this volume) that allow the corpus text and sound files to be viewed in different formats including those which are purely visual. This flexible approach is also the hallmark of several of the other projects outlined in this volume. It allows their resources to be packaged in a manner that has maximal impact for outreach given their use of state-of-the-art annotation methods and tools that capture audio, images and texts which can then be viewed in a wide range of different modes.

Naturally, it is not the case that each of the contributions to this volume exemplifies every optimal guideline we have just sketched. When taken as a whole, however, it is certainly the case that our authors demonstrate best practices regarding the key ideas introduced in this section, namely, how to treat the human subjects who contribute corpus data ethically, how to collect, digitize and manage corpus resources to effectively engage the public and how to ensure that the resources produced can be sustained longer term.

2 Outline of Contributions and Their Methodologies

The papers on creating and digitizing corpora of various kinds which made it through the review process and which we believe make excellent case studies for exemplifying the good practice for impactful corpus building sketched here fall more or less into two distinct areas of public engagement, namely, 'Education and Heritage' and 'Continuing Professional Development' (CPD). As such, the volume is organized into those chapters which largely address the issues connected with the former (Part I) and the rest which deal with corpora that have been exploited for CPD uses of different types (Part II).

Chapters 2, 3, 5 and 7 in Part I involve outreach initiatives that specifically engage pre-university (public) school audiences,⁶ while Chap. 8 offers a case study focusing on the teaching of English grammar, punctuation and spelling using smartphone apps targeted at students in schools as well as in higher education and TESOL contexts. Chapters 2, 3, 5 and 7 straddle the education/heritage divide, since they share with Chaps. 4 and 6 an orientation beyond the classroom towards the use of corpora in museum and heritage organizations as well as in more broadly defined public education projects on aspects of language and dialect awareness. Although Chaps. 10 and 11 in Part II also explore corpora and resources relevant to education, the key difference between these and the databases discussed in Part I is that the resources are not used directly with students (at least not in the first instance). Instead, they have primarily been built as professional development tools for secondary (high) schoolteachers in the case of Chap. 10 or for lecturers in higher education (Chap. 11). Chapters 9 and 12 also focus on corpora that have the potential to be exploited in workplace training though the contexts are quite different. The databases in Chap. 9 have important real-world applications for improving the knowledge base of translators and interpreters involved in Language Analysis for the Determination of Origin (LADO) procedures that feed into the legal cases of asylum seekers. Chapter 10, by contrast, reports on a set of digital resources that have applications for improving communication in everyday workplace encounters more broadly. The resources created in all of these projects, the models and methods which underpin them and their potential for generating impact in different contexts, are described more fully below.

In the first chapter of Part I, the contribution by Carolina P. Amador-Moreno, Karen Corrigan, Kevin McCafferty and Emma Moreton describes public engagement activities associated with three different but related funded projects connected by an interest in corpora associated with the Irish diaspora and/or new migrants to the region. They argue that the legacy nature of the correspondence data they have all worked with, coupled with its being generated by largely uneducated writers, presents both challenges as well as opportunities from a corpus linguist's

⁶In the North American sense of non-fee-paying, state-funded educational settings.

perspective. The chapter considers techniques for overcoming some of the drawbacks (such as the careful annotation of the corpus data using world standards for the marking up of correspondence texts). The authors demonstrate that such a strategy allows the materials to be presented in a much more user-friendly way in educational and heritage sector settings through the use of visualization techniques, for example, that would be impossible if the correspondence was not annotated in this manner.

Wendy Anderson and Carole Hough's chapter offers an impact-oriented view of various digital resources generated by public funding granted to scholars in English Language at the University of Glasgow. They focus on four distinctive corpus-building projects including the *Scottish Corpus of Texts & Speech* and its sister resource the *Corpus of Modern Scottish Writing*, as well as more recent initiatives such as *Scots Words and Place-names* and *Mapping Metaphor*. The chapter is an excellent exemplar for how best to tackle some of the issues we raised in Sect. 1.1 relating to user engagement as well as the provision of metadata and sustainability.

Sandra Clarke's contribution, which describes the development of the *Dialect Atlas of Newfoundland and Labrador* over the last 40 years, touches on similar themes. The new interactive web-based platform for the *Atlas* launched in 2013 has two main functions, that is the preservation and promotion of cultural heritage in an abstract sense in addition to providing a research and learning tool which facilitates the public's access to, and appreciation of, the distinctive regional varieties of Canada's most easterly province. The chapter also offers a model for the manner in which website resources of this kind can best be constructed so as to remain viable longer-term, a key tenet of the guidelines we put forward in Sect. 1.1.

Chapter 5 by Tyler Kendall and Walt Wolfram details what Lawson and Sayers (2016a: 3) refer to as 'the groundbreaking and multi-award-winning' engagement initiatives associated with the *North Carolina Language and Life Project*. Their contribution to this volume focuses particularly on how the public engagement spin-offs from this well-known endeavour, such as exhibitions, pedagogic tools and printed materials that include audio files delivered via CD (and most recently through advances like QR technology), interrelate with SLAAP. These initiatives all benefit from this large-scale sociolinguistic data management scheme

underpinned by the best practices argued for in Sect. 1.1 and in Kendall (2007, 2008, 2011). Their central argument here concerns the data stewardship issue introduced earlier, which they see as vital not only to ‘enrich [...] the communities from whom we obtain our data’ but also to ‘preserve that data for the study of and celebration of language variation’.

Given its orientation, many of the chapters in this volume naturally make reference to Labov’s (1982) ‘principle of debt incurred’ as well as Wolfram’s ‘principle of linguistic gratuity’ (1993, 2016; Wolfram et al. 2008) already mentioned. Bill Kretzschmar’s chapter, which follows next, openly acknowledges the debt owed to Wolfram himself with respect to the development of the *Roswell Voices* project that is its focus. The initiative was generated from local public interest in developing alongside academics a resource that would put Roswell, Georgia, on the map from a cultural heritage perspective, as Wolfram and his colleagues in North Carolina had done for communities like Ocracoke. In its present form, the *Roswell Voices* corpus consists of intergenerational interview data as well as fixed format elicitation tasks from Roswell’s African American, Latino and white (Anglo) communities so that the database captures linguistic change in apparent time in different levels of the grammar amongst diverse ethnic groups. The conversations were guided by a specific protocol which also generated local information of cultural value such as conversations about historic buildings and institutions as well as traditional farming practices. The heritage-oriented aspects of the corpus have subsequently been packaged in pamphlet and CD format in a similar manner to the exemplars described in other chapters in Part I by Amador-Moreno et al., Kendall and Wolfram as well as Mearns et al. What is unique about the *Roswell Voices* initiative, though, is its involvement in a public-private partnership attached to a European Union initiative known as The European Network of Living Laboratories. This move extended the project ‘into a more formal partnership that can address more aspects of community life, including its economic patterns’. Such a model could facilitate a novel way for corpus linguists to reach the communities with whom they wish to engage not only for data collection but also for disseminating impact.

The *Newcastle Electronic Corpus of Tyneside English* (NECTE) and the *Diachronic Electronic Corpus of Tyneside English* to which Chap. 7 by Adam Mearns, Karen Corrigan and Isabelle Buchstaller is dedicated

are very well documented (Allen et al. 2007; Beal and Corrigan 2013; Corrigan et al. 2013; Beal et al. 2014, amongst others). The open access website version of the latter, known as *The Talk of the Toon* on account of its primary function as a resource for local North East UK communities, has, however, had much less attention. The public engagement activities relating to this corpus described in the contribution to this volume by Mearns et al. were funded via a forward-thinking research council award scheme. It aimed to further develop digital research outputs like NECTE not only to ensure that they remain at the cutting edge of technological developments and are compatible with requirements for long-term sustainability, but that their accessibility should be enhanced for a wider range of audiences so as to broaden their usage and impact. Thus, as well as providing the opportunity to upgrade NECTE in various ways that improved its architecture and increased its size (and time depth), this award also gave the project team the funding for direct engagement with local schools and museums. The chapter outlines the methodology they developed for doing so as well as highlighting some of the key public outputs. Echoing the argument put forward in Amador et al. already noted, Mearns et al. also suggest that none of this would have been possible without developing the data so that it complied with world standards for corpus construction. This is a time-consuming and costly process that cannot be undertaken on the shoestring budget which some of the corpora featured here have necessarily had to deal with (like the *Roswell Voices* corpus just described, for example). Nevertheless, compliance has a number of important advantages from the perspective of the good practice guidelines featured earlier for the creation of corpora that are readily equipped to meet impact agendas.

The last contribution in Part I by Seth Mehl, Sean Wallis and Bas Aarts describes a corpus with an even longer history than NECTE/DECTE, that is the SEU which dates back to the 1950s. Its creators also share the DECTE project team's goals of developing corpora with applications for educational settings. However, a key difference in orientation between the impact tools described in Chap. 8 and the outputs arising from the corpora documented in Chaps. 1–7 is that the SEU has not so far been used as a resource for preserving or demonstrating the potential cultural heritage aspects of the data. As with all the corpora described in this volume,

however, it was developed as an academic resource initially. Nonetheless, and in keeping with the rise in demand for outreach opportunities partly generated by the UK government's edicts noted in Sect. 1.1, researchers associated with this project have turned their attention to the possibilities of using the SEU for pedagogic purposes in particular. To this end, they have developed apps for smartphones and tablets dedicated to the teaching of the English language, specifically the *interactive Grammar of English*, *Academic Writing in English* and *English Spelling and Punctuation*. One of the most important contributions that this chapter makes to the volume as a whole, and indeed to the development of the subfield termed here 'impactful corpus linguistics', is the issues which Mehl et al. raise regarding the key differences between developing a corpus resource that can be viewed as a publicly accessible website versus interacting with it on handheld devices, which present interesting challenges.

When viewed collectively, the chapters in Part I outline the manner in which the resources described (which all incorporate in some way the best practices for corpus-building advocated in Sect. 1.1) can be very successfully applied to a wide range of outreach activities. A recurrent theme is also the idea that public engagement with a research resource can best be achieved by developing these as flexibly as possible so that they can have multiple applications including those that are cultural and pedagogical.

The contributions in Part II also focus on resources that could be characterized in this way. In addition, as already noted, they are united by an interest in the manner in which professional practice can be shaped by unconventional corpora that have been built in such a way that they can generate impact of this type.

The contribution by Sjef Barbiers is a good case in point. His chapter demonstrates the manner in which a research tool originally built for the analysis of microvariation in Dutch (feeding off the corpora, databases and tools developed for the DynaSAND, GTRP and DIDDD initiatives at the Meertens Instituut) has the potential for improving the training of linguists, interpreters and translators tasked with determining for legal reasons whether or not an asylum seeker really is from the place of origin they claim (namely, implementing LADO processes). The tool in question is known as MIMORE and it includes three databases containing both

(morpho-)syntactic and (morpho-)phonological materials from a large number of locations in the Dutch language area. The key insight from the use of this tool is its value in introducing trainees to the extent to which there can be dialectal variation at different levels of the grammar so that they gain a better understanding of how regional differences can be marked linguistically. It is also useful in demonstrating the complications that arise when using tools like this to locate individual speakers with absolute confidence.

Chapter 10 by Jenny Cheshire and Sue Fox begins with a discussion of the educational context which prompted the development of the CPD outputs tied to their London corpus and then broadens out into a description of the resource and its value for outreach activities relating to teacher training. The two projects known as *Linguistic Innovators* and *Multicultural London English* which took place between 2004 and 2010 resulted in a corpus of nearly 3.5 million words. The goals of the research included testing the Londoncentric nature of linguistic change across the UK and investigating processes relating to the acquisition and spread of innovations in dynamic, multilingual metropolitan contexts. Follow-on funding then allowed them to set up spin-off projects. These permitted the repurposing of the corpus data to create a Teaching Resources Online Archive. It is geared particularly to final-year secondary (high) school examinations in English Language and is closely tied to the Advanced level curriculum so that the Archive is maximally useful to the staff delivering these courses. One of the most important aspects of this contribution which we wanted to highlight here is their discussion of the ethical issues that arise when academic resources are repurposed for wider engagement initiatives. As Cheshire and Fox rightly point out here, the anonymization protocol required for ethical reuse of this kind of corpus (especially where audio data is concerned) needs to be especially stringent when the materials are intended for applications in the public domain. They demonstrate the challenges that they faced when developing their own CPD resources and lay out a set of best practices of the kind already alluded to in Sect. 1.1.

The penultimate chapter in the collection by Steve Walsh and Dawn Knight also concerns a corpus originally collected for academic purposes that has tremendous potential for the training of teachers—though it is

in a higher education context this time. The *Newcastle University Corpus of Academic Spoken English* (NUCASE) is a corpus of 1 million words recorded in small group contexts across different university disciplines at Newcastle University which ‘provides a “snapshot” of spoken academic discourse in contexts where interaction takes place’. The chapter focuses on how insights from both corpus linguistics and conversation analysis can better inform our understanding of the ways in which ‘interactants establish understandings in educational settings and, in particular, highlights the inter-dependency of words, utterances and text in the co-construction of meaning’. Having established their analytical model, Walsh and Knight then go on to demonstrate techniques by which NUCASE might be exploited in CPD activities for university lecturing staff as well as those designing materials for use in small-group teaching so as to improve the learning experience associated with this mode of delivery in higher education.

The chapter by Bernadette Vine also examines language in workplace settings but these are far beyond academia (including building sites and hospital wards). Moreover, the Wellington *Language in the Workplace* corpus contains participants who have differing types of role relationship than the lecturer versus student hierarchy represented in Walsh and Knight’s corpus. What is more, the data is drawn from a much more diverse range of interactional exchanges, from the briefest comments on busy and noisy factory lines to more extended tea-break conversations. A key feature of the engagement associated with this endeavour (as indeed with many of the other projects articulated in this volume) is that knowledge is thought of as ‘exchanged’ in a two-way process rather than as ‘transferred’ with the implication of a hierarchy from ‘gown’ to ‘town’. Vine articulates this stance in her chapter in the following manner: ‘engagement with workplaces has helped shape our research objectives and understanding, as well as assisting research participants to develop a fuller appreciation of the importance of effective communication’. Great efforts have also been made to raise awareness of the results of the research amongst the wider business community in Wellington and further afield by publicizing it in newsletters, non-academic journals and the mass media. The team has also used the corpus to develop teaching materials and resources (not dissimilar to the Archive created by Cheshire

and Fox) with the aim of assisting both native and non-native speakers to become more effective workplace communicators.

The chapters in Part II highlight the practical applications of unconventional corpora for CPD in a range of professions. It is interesting to note that, with the exception of the contribution by Vine in which the data was co-designed by academics and end-users, these outputs have actually been serendipitous and the resources have come to be applied in contexts that the researchers are unlikely to have considered at the time of grant application. This again highlights the value of building a flexible corpus. It also underlines the significance of making the most of what Lawson and Sayers (2016b: 21) in their discussion of impactful sociolinguistics describe as ‘the importance of chance encounters’ and the ‘seizing of opportunities when they arise’.

3 Acknowledgements

The editors would like to close by acknowledging the financial support provided to various phases of this project by the Arts and Humanities Research Council (grant nos RE11776, AH/H037691/1 and AH/K008285/1).

We would also like to express our deeply felt gratitude to our authors who have gracefully endured our cajoling to shape both their content and format. They have also actively engaged with us in pursuing a research agenda in corpus linguistics with a very specific orientation, namely, one in which the commitment to ‘paying our debts to the communities whose data have helped to build and advance our careers’ (Rickford 1993: 130) is taken very seriously. In addition to articulating international standards for metadata, and best practices for the collection, preservation, and annotation of corpus data as recommended by Kretzschmar et al. (2006), these scholars also orientate towards the creation and digitization of corpora in a manner which makes them highly suited to engaging non-academic audiences.

There are also a number of other people who deserve special thanks, including: Esme Chapman, Chloe Fitzsimmons, Elizabeth Forrest, and Olivia Middleton, the Palgrave Macmillan editors responsible for this

companion volume, for their helpful feedback from inception to completion; other staff at Palgrave Macmillan for their patience with our many technical queries; Claire Childs and Christine Wallis for their assistance with formatting and indexing; the organizing and scientific committees of the *Methods in Dialectology XIV* conference at the University of Western Ontario and *Corpora Galore 2011* at Newcastle University. Thanks are also due of course to our anonymous reviewers who submitted the conference papers and the chapters in this volume to critical scrutiny. Finally, we are indebted to Joan Beal for writing the foreword to this volume and for the many discussions we have had with her regarding the shape that the book should take since the idea for this project was first mooted back in 2011. Any remaining shortcomings are, as usual, our own.

References

Books and Articles

- Allen, Will, Joan C. Beal, Karen P. Corrigan, Hermann L. Moisl, and Warren Maguire. 2007. The *Newcastle Electronic Corpus of Tyneside English*. In *Creating and Digitizing Language Corpora: Vol. 2, Diachronic Databases*, eds. Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, 16–48. Basingstoke: Palgrave Macmillan.
- Bauer, Laurie. 2004. Inferring variation and change from public corpora. In *The Handbook of Language Variation and Change*, 1 edn, eds. J.K. Chambers, and Natalie Schilling, 97–114. Malden: Blackwell.
- Beal, Joan C., Karen P. Corrigan, and Hermann L. Moisl, eds. 2007a. *Creating and Digitizing Language Corpora: Vol. 1, Synchronic Databases*. Basingstoke: Palgrave Macmillan.
- , eds. 2007b. *Creating and Digitizing Language Corpora: Vol. 2, Diachronic Databases*. Basingstoke: Palgrave Macmillan.
- Beal, Joan C., and Karen P. Corrigan. 2013. Working with unconventional existing data resources. In *Data Collection in Sociolinguistics: Methods and Applications*, eds. Becky Childs, Christine Mallinson, and Gerard van Herk, 213–216. London: Routledge.
- Beal, Joan C., Karen P. Corrigan, Adam J. Mearns, and Hermann L. Moisl. 2014. The *Diachronic Electronic Corpus of Tyneside English*: annotation and dissemination practices. In *The Oxford Handbook of Corpus Phonology*, eds.

- Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 517–533. Oxford: Oxford University Press.
- Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, Ben Rampton, and Kay Richardson. 1997. Ethics, advocacy and empowerment in researching language. In *Sociolinguistics*, eds. Nikolas Coupland, and Adam Jaworski, 145–162. Houndmills: Macmillan. (Originally published in *Language and Communication* 13(2): 81–94 in 1993.)
- Childs, Becky, Gerard van Herk, and Jennifer Thorburn. 2011. Safe harbour: ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory* 7(1): 163–180.
- Choudrie, Jyoti, Susan Grey, and Nicholas Tsitsianis. 2010. Evaluating the digital divide: the Silver Surfer's perspective. *Electronic Government, An International Journal* 7(2): 148–167.
- Corrigan, Karen P., Adam J. Mearns, and Hermann L. Moisl. 2013. Data-mining the DECTE Corpus: phonological and morphological variability in Tyneside English. In *Cross-Linguistic and Language-Internal Variation in Text and Speech*, eds. Benedikt Szmrecsanyi, and Bernhard Wälchli, 113–149. Berlin: Walter de Gruyter.
- D'Arcy, Alexandra. 2011. Corpora: capturing language in use. In *Analysing Variation in English*, eds. Warren Maguire, and April McMahon, 49–71. Cambridge: Cambridge University Press.
- Day, Timothy. 2001. The National Sound Archive: the first fifty years. In *Aural History: Essays on Recorded Sound*, ed. Andy Linehan, 41–64. London: The British Library.
- Durand, Jacques, Ulrike Gut, and Gjert Kristoffersen, eds. 2014. *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press.
- Kendall, Tyler. 2007. The Sociolinguistic Archive and Analysis Project: empowering the sociolinguistic archive. *Penn Working Papers in Linguistics* 13(2): 15–26.
- . 2008. On the history and future of sociolinguistic data. *Language and Linguistics Compass* 2(2): 332–351.
- . 2011. Corpora from a sociolinguistic perspective. *Revista Brasileira de Linguística Aplicada* 11(2): 361–389.
- Kretzschmar, William A., Jean Anderson, Joan C. Beal, Karen P. Corrigan, Lisa-Lena Opas-Hänninen, and Bartek Plichta. 2006. Collaboration on corpora for regional and social analysis. *Journal of English Linguistics* 34(3): 172–205.
- Labov, William. 1982. Objectivity and commitment in linguistic science. *Language in Society* 11: 165–201.

- Lawson, Robert, and Dave Sayers. 2016a. Introduction. In *Sociolinguistic Research: Application and Impact*, eds. Robert Lawson, and Dave Sayers, 1-6. London: Routledge.
- Lawson, Robert, and Dave Sayers. 2016b. Where we're going, we don't need roads: the past, present, and future of impact. In *Sociolinguistic Research: Application and Impact*, eds. Robert Lawson, and Dave Sayers, 7-22. London: Routledge.
- Martin, Ben R. 2011. The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster? *Research Evaluation* 20(3): 247–254.
- Norris, Pippa. 2001. *Digital Divide: Civic Engagement, Information Poverty and the Internet in Democratic Societies*. New York: Cambridge University Press.
- Perks, Robert P. 2011. Messiah with a microphone? Oral historians, technologies and sound archives. In *The Oxford Handbook of Oral History*, ed. Donald A. Ritchie, 315–332. Oxford: Oxford University Press.
- Reaser, Jeffrey, and Carolyn Temple Adger. 2007. Developing language awareness materials for non-linguists: lessons learned from the Do You Speak American? project. *Language and Linguistics Compass* 1(3): 155–167.
- Rickford, John. 1993. Comments on 'ethics, advocacy and empowerment'. *Language and Communication* 13(2): 129–131.
- Robertson, Beth M. 2011. The archival imperative: can oral history survive the funding crisis in archival institutions? In *The Oxford Handbook of Oral History*, ed. Donald A. Ritchie, 393–408. Oxford: Oxford University Press.
- Rowlands, Ian, David Nicholas, Peter Williams, Paul Huntington, Maggie Fieldhouse, Barrie Gunter, Richard Withey, Hamid R. Jamali, Tom Dobrowolski, and Carol Tenopir. 2008. The Google Generation: the information behaviour of the researcher of the future. *Aslib Proceedings* 60(4): 290–310.
- Samuel, Gabrielle N., and Gemma E. Derrick. 2015. Societal impact evaluation: exploring evaluator perceptions of the characterization of impact under the REF2014. *Research Evaluation* 24: 229–241.
- Smith, Abby, David Allen, and Karen Allen. 2004. *Survey of the State of Audio Collections in Academic Libraries*. Washington, DC: Council on Library and Information Resources.
- Wolfram, Walt. 1993. Ethical considerations in language awareness programs. *Issues in Applied Linguistics* 4: 225–255.
- . 2012. In the profession: connecting with the public. *Journal of English Linguistics* 40(1): 111–117.

- . 2013. Community, commitment and responsibility. In *The Handbook of Language Variation and Change*, eds. J. K. Chambers and Natalie Schilling. 555-576, 2. Malden: Wiley/Blackwell.
- . 2016. Public sociolinguistic education in the United States: a proactive, comprehensive program. In *Sociolinguistic Research: Application and Impact*, eds. Robert Lawson, and Dave Sayers. 87-108. London: Routledge.
- Wolfram, Walt, Jeffrey Reaser, and Charlotte Vaughan. 2008. Operationalizing linguistic gratuity: from principle to practice. *Language and Linguistics Compass* 2(6): 1109–1134.

Websites and Online Resources

- RCUK Policy on Open Access*. <http://www.rcuk.ac.uk/research/openaccess> (accessed 24 June 2015).
- REF: Research Excellence Framework*. <http://www.ref.ac.uk/pubs/2011-02> (accessed 24 June 2015).
- W3C: World Wide Web Consortium*. <http://www.w3.org> (accessed 13 July 2015).

Part I

Corpora for Education and Heritage

2

Migration Databases as Impact Tools in the Education and Heritage Sectors

Carolina P. Amador-Moreno, Karen P. Corrigan,
Kevin McCafferty, and Emma Moreton

1 Introduction

There has been considerable recent investment in the digitization of databases, like the *Documenting Ireland: Parliament, People and Migration* (DIPPAM) project, that relate in various ways to the history and dias-

The authors are grateful for funding received from the UK's Arts and Humanities Research Council (AHRC) in connection with the DEM project (grant no. AH/K006231/1) and *Múin Béarla do na Leanbháin* (MBDNL) ('Teach the Children English') project (grant no. AH/K008285/1). CORIECOR was funded by the University of Bergen's Meltzer Foundation and the Research Council of Norway (grant no. 213245) and we would also like to acknowledge their support. We similarly wish to thank our anonymous reviewers for extremely helpful comments on an earlier draft.

C. Amador-Moreno (✉)
University of Extremadura, Cáceres, Spain

K.P. Corrigan
Newcastle University, Newcastle upon Tyne, UK

K. McCafferty
University of Bergen, Bergen, Norway

E. Moreton
Coventry University, Coventry, UK

© The Editor(s) (if applicable) and The Author(s) 2016
K.P. Corrigan, A. Mearns (eds.), *Creating and Digitizing
Language Corpora*, DOI 10.1057/978-1-137-38645-8_2

pora of Ireland, which has been an area of intensive scholarship since the later twentieth century (see, for example, Miller 1985; O'Sullivan 1992a, b, c, d, e, f; Fitzgerald and Lambkin 2008; Miller 2008). As such resources were largely designed for academics in historical studies and allied disciplines, their applicability as tools to engage public audiences (particularly in the education and heritage sectors) remains to be tested. This chapter discusses best practices in the creation of databases like these so they can be exploited for a much wider variety of academic and non-academic uses by focusing on two related digital initiatives, namely, the *Corpus of Irish English Correspondence* (CORIECOR) project currently being undertaken at the University of Bergen and Coventry University's *Digitizing Experiences of Migration: The Development of Interconnected Letter Collections* (DEM) project completed in 2014.

CORIECOR is a collection of emigrant writings incorporating some of the letters from DIPPAM (largely eighteenth- to twentieth-century data) as well as an Irish-Argentinian collection (nineteenth century) (Amador-Moreno and McCafferty 2012).

The DEM project was an AHRC research network project funded under their 'Digital Transformations in the Arts and Humanities' theme. A key objective of the scheme was to fund research that would 'exploit the potential of digital technologies to transform research in the arts and humanities, and to ensure that it is at the forefront of tackling crucial issues such as intellectual property, cultural memory and identity, and communication and creativity in a digital age'. DEM's specific remit was to explore the digitization and annotation of historical emigrant letter collections, with a special focus, as CORIECOR also has, on letters of the Irish diaspora.

This chapter offers case studies of best practice in extending the utility of the data sets upon which the CORIECOR and DEM initiatives are based beyond historical studies so that they become useful for other disciplines and for aspects of primary and secondary teaching as well as informing public lectures and the like which contribute to lifelong learning. A model for utilizing these corpora in the heritage industries is also presented which demonstrates how they can be exploited by this sector to fulfil its key goal of preserving and interpreting cultural heritage.

The chapter begins with an outline of the contents of the primary data sets and the manner in which they have been digitized. It also includes a discussion of correspondence corpora as primary data suitable for

sociolinguistic analysis, as well as an outline of the study of letters as cultural artefacts since both of these uses can be incorporated in public engagement activities. This is then followed by an exploration of how the CORIECOR corpus—despite ultimately deriving from a historical migration database like DIPPAM—can, nevertheless, be very successfully mined to further our understanding of linguistic change. The report of the DEM project, which follows, focuses particularly on the issues that arise when attempting to digitize correspondence corpora like CORIECOR using techniques which comply with world standards for the encoding of digital text. Finding a mechanism for doing so is, however, key not only to their sustainability longer term and their interoperability but also to the ease with which they can be exploited in educational and museum settings.¹ The chapter then closes with an outline of some examples of how these data sets have been used for such public engagement purposes.

2 Irish Migration Databases

The collaborative *Documenting Ireland: Parliament, People and Migration* (DIPPAM) project is an online archive of sources hosted by Queen's University, Belfast, that relates to the history of Ireland and the migratory experiences of its population between 1700 and the twentieth century.² It consists of three principal databases: (a) *Enhanced British Parliamentary Papers on Ireland* (EPPI); (b) *Irish Emigration Database* (IED)³ and (c) *Voices of Migration and Return* (VMR). Key to the DIPPAM project was the development of interoperability across each database in order to facilitate cross-searching, allowing users to seek out keywords simultaneously in EPPI, IED and VMR. Collectively, the website includes materials relating to: (a) correspondence and family memoirs conveying personal

¹By 'interoperability' we mean the possibility that information can be exchanged seamlessly between one database and another. To achieve this, certain standards need to be adhered to when building the corpora in the first place and we elaborate further on these in Sect. 4.

²The project is the result of collaboration between Queen's University Belfast, the Mellon Centre for Migration Studies (MCMS) in Omagh, the University of Ulster and Libraries NI, and was also funded by the AHRC.

³This subcomponent is key to the discussion in this chapter as it is a collection of over 4000 letters by Irish emigrants, held in various archives including the Public Record Office of Northern Ireland (PRONI). It formed the basis of CORIECOR when the project first began.

The top screenshot shows the IED (Irish Emigration Database) interface. The main content area displays a newspaper advertisement for the Columbia et al. ships, dated 15th March, 1825. The text describes the ships' departure for Baltimore and lists the captain and passengers. A metadata table on the right provides details such as Document ID (9802411), Date (23-02-1825), Document Type (Newspapers), and Citation (Ships Columbia et al., Belfast to U.S. & Canada).

The bottom screenshot shows the EPPI (Enhanced British Parliamentary Papers on Ireland) interface. The main content area displays a bill titled "Bill for Protection and Relief of Destitute Poor evicted from Dwellings in Ireland (as amended by Lords)". A "Document Metadata" table is visible, listing details such as Source (HC), Paper No (470), LC Subject Heading (Ireland - History - Famine, 1845-52/Poor laws - Ireland/Public welfare - Ireland), Breviate Keywords (Irish papers - distress arising from the Famine), Publisher (HMSO), Breviate Page (403), and Series (Sessional papers).

Fig. 2.1 An advertisement for the Columbia et al. ships, Belfast to USA and Canada (*top*), and a bill for protection and relief of destitute poor evicted from dwellings in Ireland (*bottom*)

narratives of experience focusing on both historical and more recent migration to and from Ireland; (b) diaries, journals, wills and newspaper extracts (including family announcements as well as shipping news and advertisements such as that in Fig. 2.1); and (c) parliamentary papers documenting the social context of Irish migration from 1800 to 1922. Particular emphasis in the selection is placed on sessions that relate to key sociohistorical events in Ireland like the Great Famine (see Fig. 2.1), which precipitated the unprecedented demographic dislocation examined in Kennedy et al. (1999) and Ó Gráda (2000) *inter alia*.

This collection of fully searchable and browsable text documents is multimodal (see Knight 2011) in that DIPPAM also contains images and audio files, which have been used for example as evidence for the analyses presented in Devlin Trew (2013).⁴ As noted in the project's entry on the Research Councils UK 'Gateway to Research' website, DIPPAM was created for facilitating and encouraging 'new kinds of research into modern Irish history through juxtaposing and integrating a variety of sources—quantitative and qualitative, official and unofficial, public and private'.⁵ Although it was never specifically designed to facilitate research outside of historical studies and allied disciplines, the IED subcorpus of DIPPAM, has, nevertheless, become the backbone of CORIECOR which has been used very successfully to explore linguistic variation and change in Irish-English (IrE), as we demonstrate in Sect. 3 below.

2.1 The Linguistic and Sociocultural Value of Irish Migration Databases

Letters can be considered part of what constitutes the notion *intrahistoria*, a term coined by the Spanish writer Miguel de Unamuno in 1895, and 'rescued' by the historian José María Jover. It refers to the value of the humble and anonymous lives experienced by ordinary men and women in everyday contexts which form the essence of normal social interactions, as opposed to the lives of leaders and famous people that are generally accounted for in canonical histories (Earle 1964; Valdés 1996). Personal correspondence offers a rich and colourful view into the past and has long been used as a valuable source for writing what Elspaß (2005) has called 'language history from below' (*Sprachgeschichte von unten*). In the context of Irish emigration, historians such as those who compiled DIPPAM have used personal letters written by merchants, farmers, peasants, artisans and labourers for more than 50 years.⁶ Emigrant letters

⁴ See, for instance, <http://www.dippam.ac.uk/vmr/interview/463>, an interview with a 60-year-old man from County Donegal.

⁵ See <http://gtr.rcuk.ac.uk/project/12B979D5-D08B-4B5B-92A5-58B5B292A587>.

⁶ See, for example, Arnold Schrier's groundbreaking *Ireland and the American Emigration 1850-1900* (first published in 1958).

can give direct access to the thoughts, feelings, ambitions, ideologies and economic motives that over the last four centuries have led more than 11 million Irish to leave their country for Argentina, the Antipodes, Great Britain, North America, South Africa and other far-flung places. Their letters can tell us a lot about emigrants' relations with one another, with Irish emigrant communities and with the wider societies in the countries they migrated to, as well as the families and communities they left behind in Ireland. They thus give important insights into the push/pull factors of migratory processes, the role that institutions and communities play in supporting emigration and they also offer a unique perspective on how individuals and groups adapted to the New World (see Thomas and Znaniecki 1958; Erickson 1972; Miller 1985; Kamphoefner et al. 1988; Fitzpatrick 1994).

Correspondence corpora also transcend disciplinary and methodological boundaries. Hence, Michael Montgomery, who pioneered the use of personal letters for the linguistic study of Ulster Scots, has noted that, although linguists had until the mid-1990s paid little attention to this type of material, such texts likewise provide access to the language of the common people (Montgomery 1995: 28). Correspondence of this kind has, however, been regarded as problematic from a historian's point of view, as Elliott et al. (2006: 3–4) point out:

There is also the problem of making sense of the writings of the many poorly educated and marginally literate writers for whom the necessity of writing inspired an exercise that taxed their abilities to the limit. Decoding texts with inadequate paragraphing and punctuation, ungrammatical constructions, highly irregular spelling, and language that combines native regional dialect, borrowings from the tongue of the country of resettlement, archaic colloquialisms, and singular, individualized modes of expression is the common experience of those who read immigrant letters.

However, the historian's difficulty can also be seen as the linguist's opportunity (McCafferty 2016), given that it is precisely the ungrammatical constructions, the irregular spelling, the native regional dialect, the archaic colloquialisms, and so on—engendered by the low literacy levels

of these periods—that provide us with evidence for linguistic analysis.⁷ This type of writing represents what the Portuguese philologist José Leite de Vasconcelos referred to as ‘textos errados’, which, as he pointed out in 1890, deserved as much attention as educated speech:

Quem possui pouca cultura litteraria escreve muitas vezes como falla, não só por ignorar frequentemente as regras grammaticas, como porque nessas pessoas tem mais fôrça o habito da pronuncia do que o da escrita; ora os erros então cometidos, erros, já se vê, em relação ás normas preestabelecidas servem para o lingüista, porque lhe revelam exactamente o que ella procura. [...] Toda lingua propriamente dita, quer seja popular, quer culta [...] quer nela estejam esculpidas as epopeias homéricas, quer sirva só para as limitadas relações sociaes de um canto de provincia—é uma lingua perfeita, uma lingua que merece as attenções da sciencia, porque representa a verdade. (Leite de Vasconcelos 1890: 16–17)

[[Those] who have little literary culture often write as they speak, not only because they ignore grammatical rules, but because for them pronunciation is a stronger habit than spelling, they are more used to speaking than to writing. The errors they make (called *errors*, of course, only in relation to the standards of established norms) are useful for the linguist because they reveal exactly what he is looking for. [...] Everything we can call language, be it popular language or learned language, either engraved with the Homeric epic or serving the limited social relations of one corner of the country, is a perfect language, a language that deserves the attention of science, because it is the truth].⁸

In the context of emigrant letter writing, the spontaneity of the speech of the less educated members of society is also significant, given that they were ‘the more likely to have speech intrude into their writing because their limited literacy made them more dependent on their ear’ (Montgomery 2001: 13).

Like conversation, letters are interactive, and therefore, imply participation. There is a recipient, a sender (or encoder), and a message

⁷For further details regarding the particular case of Ireland, see the arguments in Daly (1990) and Corrigan (2003).

⁸Translation by Carolina P. Amador-Moreno.

that needs to be decoded in a particular context. As Dossena (2012: 28) has noted:

the attempt to make the message accessible to somebody else is a precondition for writing, and all encoders strive to ensure that what is conveyed will be decoded appropriately. Recipients, whether a few hours, days, months or even centuries away, are already in the world of the letter. Nor can they be out of it: their presence, albeit virtual, is what makes the text meaningful, and every choice is both dictated by and helps maintain a continuing relationship between participants.

In a sense they are not too different from other types of communication in the modern world, where the rise of multimodal frameworks and the resurgence of written modes (which reproduce even more closely the features of conversational language) seem to call for a refocus on writing as a legitimate object of sociolinguistic analysis (Lillis 2013: 2–3). Computer-mediated communication in the present day context (for example, emails, chat rooms, blogs, social media, WhatsApp, SMS, and so on) have come to replace the less immediate communicative practice of private correspondence, and yet they share with this old mode of writing letters their personal, unselfconscious and spontaneous nature. Needless to say, there can be no assumption that this type of data is not edited (or even re-edited) by the author of the message, but spoken language is not immune from these traits either, since it often displays repetition, reformulation, false starts and explanatory clarifications. Excerpt (1) from a CORIECOR letter written in 1889 by John Stevenson Sinclair is a good case in point. He writes before an upcoming visit to Ireland and stresses that he is not planning to spend the night in Sixtowns, where his family and friends live, but instead intends to stay at a hotel in a town some distance away. The kind of repetition that he employs to convey his message would not be characteristic of the writing of practised writers, but it can often be found in informal conversation. Indeed, repetition of just this kind is an important part of the normal run of speech:

- (1) I may have Miss
Margaret Sinclair a long with
me she says she will stop

with her Aunt Margaret and
 I will stop in Cookestown Hotel
 Cal John is goying to be verry
 Toney in Ireland or I wood not
 go I will go by Tolybrick,
 I will not let Marey know in
 Sixtowns unto I drive town the
 Tolybrick Hill I will leave Margaret
 at your House but I will not stop
 I will Drive around by the old
 Church and back to Cookstown
 I don't Intend to sleep a night
 in Sixtowns I will visit 3 Houses
 in Sixtowns regular but I will
 make my home in Cookstown Hotel
 you mea know the reason I have
 for not stopping in Sixtown Margaret
 I gess you doe but I tell you
 before I go I will not stop
 in Sixtowns I wish you all well
 every 1 of my cousins P Heron
 and your House but I will not
 stop

This letter by Stevenson Sinclair illustrates the orality of private correspondence, which can be observed also in the use of non-standard forms, dialect features and colloquial language in general. The system of modality, for example, is known to have undergone radical changes in the history of English. Modals (for instance *must*) have decreased in use, while semi-modals (for example *have to*) and stance adverbs, which convey the speaker's/writer's attitude (such as *certainly* versus *seemingly*), have, in general, undergone an increase historically, particularly in personal, colloquial registers such as private letters. This phenomenon leads Biber and Conrad (2009: 173) to the conclusion that there has been a general shift in cultural norms: 'speakers and writers are more willing to express their personal attitudes and evaluations in recent periods than in earlier historical periods' (see also, for example, Krug 2000; Leech 2003; Close and Aarts 2010; Fehringer and Corrigan 2015).

The kind of diachronic development undergone by semi-modals and stance adverbials is termed *colloquialization* by some linguists (Biber 2003; Mair 2006), who observe that the English language has become more informal, even more speech-like, over the last two centuries or so. Examples of exactly this phenomenon in CORIECOR letters can be illustrated by the use of the discourse marker *like* in clause-final position. This has been documented as a discourse feature of older British and Irish dialects (Tagliamonte 2012: 167) as well as contemporary ones, as in example (2) from an interview with a young teenager of Polish extraction now residing in NI, conducted in 2014 for the Múin Béarla do na Leanbháin project (MBDNL, see Sects. 5.2 and 5.3). Extract (3), which is remarkably similar despite the passage of time, is from John Stevenson Sinclair's letter again. (4) is from another CORIECOR letter—this one having been sent by David McCullough from New Zealand to his family in Co. Down:

- (2) She has the opportunity to socialise with other people like
(Extract from the MBDNL corpus interview with 'Polly Wiczorek',
Armagh, dated 2014)⁹
- (3) I will Have Ø return ticket I have to see Robert like
(JSS, 05.03.1889)
- (4) the [they?] have to work for nothing that month like
(DMcC, 04.06.1876)

Both tokens in (3) and (4) show that IrE speakers used this type of discourse marker by the end of the nineteenth century, and the fact that clause-final *like* is documented in letters should alert us to look for its occurrence in other similar texts (see Amador-Moreno and McCafferty, 2016). At the same time, it is an indication of the orality and informality of the written sources in which they appear since the contemporary example in (2) was digitally recorded in just those circumstances.

Sound recording equipment has, however, only been around for about a hundred years, so for periods before the early twentieth century, we have to use written texts like CORIECOR if we want to investigate

⁹All the names in the MBDNL corpus have been changed to ethnically appropriate pseudonyms so as to preserve the participants' anonymity.

language change, since writing was the only way of capturing and storing language until the advent of recording technology.

Anyone interested in IrE might record people in conversation as the MBDNL researchers have done and study the language of the recordings in minute detail to show not only how individuals use the English language in a certain place and at a particular point in time, but also how usage patterns are distributed across communities throughout the island of Ireland. If we repeat this procedure at regular intervals, say every 20 years, the shifting patterns of language use will eventually reveal how the English of the Irish changes through time. If we had been able to do this every two decades or so since the Plantation era, we would have a priceless sound archive that could be used by linguists to trace the evolution of IrE out of the meeting of dialects of English and Scots, with an admixture of Irish and Scottish Gaelic. Unfortunately, there is no such recording archive, but private correspondence written by people with only rudimentary education, as discussed above, can provide interesting insights into the speech of those who communicated with friends and relations through letters, and who often wrote because they had to if they wanted to keep in touch with those who had been left behind. Until the twentieth century, a written letter or postcard was the only means of contacting one another across large distances. These were written by people who had no expectation that their writings would ever be published or collected in archives, but it is in exactly this type of material that we find evidence for the more colloquial, dialectal or vernacular language of the past. We elaborate on this point in Sect. 3 below.

3 Building and Analysing CORIECOR

CORIECOR makes available a large body of letters, written by Irish people of both genders and all social and geographical backgrounds, sent to and from Ireland from about 1700 to 1940. It thus provides an excellent empirical base for studies of historical change in IrE. CORIECOR is currently under development. At the time of writing (May 2015), it consists of 4793 personal letters (over 3 million words) written to and from Irish emigrants from the 1670s onwards. Coverage is good from

the 1760s to the 1940s (minimum 55,000 words per 20-year subperiod). As already noted, most of the texts come from the DIPPAM collection. The northern province of Ulster and the east coast region of Leinster are over-represented in CORIECOR, especially in the earlier subperiods, but this bias is, in part, a reflection of the fact that these were the regions where English was widely spoken before 1800. Several new collections found in Argentina, Australia, Canada and the USA have recently been added, which provide data from under-represented subperiods (like the early eighteenth century) and improve the geographical spread of the material. Of the texts available, 4828 letters have already been classified by individual informants (letter writers in this case). This will aid historical sociolinguistic study, enabling us to extract and encode data, taking account of aspects of the backgrounds of individual writers (geographical origin, gender, social status, ethnicity/religion, social mobility, and so on) and other factors that might influence language use (relationship to addressee, purpose of writing, for example). Individuals have also been grouped into networks of family, friends, colleagues, business associates and the like, to facilitate research based on social network approaches that have proved fruitful in the study of the present-day language (Milroy 1987) and earlier English (Fitzmaurice 2007).

One of the challenges of sociohistorical corpus research based on texts like the CORIECOR letters is that it builds on a kind of empirical data that is not easy to process automatically.¹⁰ However, one of the goals of the CORIECOR project was to store the texts as TEI-conformant XML documents, in accordance with the standards recommended by the World Wide Web Consortium (W3C).¹¹ The corpus will also eventually be made available for download in XML format as well as in plain text versions which are more suitable for wider audiences, as advocated by Mearns et al. (this volume).

The search capabilities of, for instance, *Corpus Presenter 14* (Hickey 2014) and *Wordsmith Tools 6.0* (Scott 2012) have been used to extract

¹⁰ For a discussion of automatic processing methods see Hendrickx et al. (2011).

¹¹ See <http://www.w3.org/standards/xml>. TEI refers to a consortium which collectively develops and maintains a standard for representing texts in digital form. XML stands for 'extensible mark-up language' and it is used to add annotations or additional information which are not visible to the end-user but are needed by the computer so that the text can be read and processed correctly.

data for CORIECOR-based academic purposes so far. Both packages offer rapid search and text manipulation functions but the latter is much more widely used than the former. Although both should be relatively easy to learn and apply in secondary-school contexts, they each have two key limitations which makes their applications outside higher education somewhat problematic. Neither of these software packages is free, for example, nor are they compatible across all platforms since they are designed with PC users in mind rather than Apple Macintosh or Linux/Unix.¹² Moreover, the rise of iPad/tablet usage in teaching and learning contexts outside of higher education will no doubt also have an impact on their applicability in school settings.¹³

As such, school users who have never worked with corpus analytic tools before or who want to use CORIECOR for less complex searches might well find *AntConc* (Anthony 2015) more accessible. It has the major advantage of being free and fully operational on PC, Mac and Unix/Linux operating systems and there are also clear and simple tutorials available on how to make the most of its functions.¹⁴

As far as academic end-users are concerned, the *Goldvarb X* for Windows/*Goldvarb Lion* for Macintosh statistical packages (Sankoff et al. 2012), specially developed for research on linguistic variation, have also been very successfully used by members of the CORIECOR team as a tool for managing extracted data, and preparing and performing statistical analyses of variation, using the methodology advocated by Tagliamonte (2006). This approach has been developed specifically for use with large corpora designed to study language variation and change. However, the outcomes of analyses using this tool and other statistical packages like it have already proven useful in the contexts of education and the heritage industries in order to present data to wider publics, as Mearns et al. (this volume) show. By publicizing the results of linguistic analyses in this manner, issues of language, identity and integration/alienation can be explored using the medium of museum exhibitions and the like, as we demonstrate in Sect. 5 below.

¹² *Wordsmith* can, however, be used by Mac users who have access to PC emulator software like Parallels (see: <http://www.parallels.com>), though that is not freeware either.

¹³ See NAACE: <http://www.naace.co.uk/publications/longfieldipadresearch>.

¹⁴ See the Video Tutorials links in the User Support section on the *AntConc* website: <http://www.laurenceanthony.net/software/antcon>.

3.1 Investigating Irish English Features across Time

CORIECOR gathers as much evidence as possible for early IrE into a corpus that permits long-term diachronic study, which will for the first time allow researchers to trace the emergence and development of features of this variety, including stylistic, regional and social variation. The corpus can be utilized for empirical comparisons of IrE with data from other sources for the Late Modern English period (1700–1945). Apart from the discourse markers introduced in Sect. 2.1 (see also Amador-Moreno and McCafferty 2015), the CORIECOR researchers have also been investigating diachronic change in phonology and grammar.

The use of first-person *will* versus *shall*, for example, is interesting because it is historically one of the most widely commented-on ways in which IrE deviates from standard English (Tieken-Boon van Ostade 2009: 90–1). The preference for first-person *will* was condemned by normative grammarians throughout the eighteenth and nineteenth centuries as an Irish (and Scottish) *misuse* (McCafferty and Amador-Moreno 2012: 185–6). However, although the recent shift towards *will* with all persons in North American English—now also affecting British English—has been attributed to the influence of Irish immigrants, evidence from the letters (McCafferty and Amador-Moreno 2014) shows that while IrE shifted rapidly towards *will* by the 1880s, it was not unusual in this respect. In fact, a similar development took place at the same time in Canadian English, which may indicate a more general trend, at least in colonial Englishes. CORIECOR thus reveals that there are good reasons to doubt that it was the influence of IrE which actually drove the change towards first-person *will*.

A number of phonological aspects of IrE are also reflected in the CORIECOR letters. These are typically instances where the letter writers are spelling words phonetically, within the constraints of the normal Latin alphabet, as a representation of the way they hear them pronounced in the spoken language. Given that, as we argued in Sect. 2.1, many of the letter writers must have had only the most rudimentary education, it is not surprising that the orthography of even some of the most basic vocabulary in the language is found in spellings reflecting vernacular pronunciations. This is the case, as de Rijke ([forthcoming](#)) argues, with words

such as *Henery* or *worald*, which show schwa-epenthesis, a well-known feature of present-day IrE that occurs across the island, most typically in liquid–nasal environments (for instance, *film* and *farm*). In his study of the representation of epenthesis in letters written over a period of more than 200 years, de Rijke documents possible clusters containing the most common words affected by this phonological feature diachronically. His study shows that epenthesis is well attested in the corpus and is found in a wider range of clusters (for example, /wn/, /dr/, /ŋr/, /fl/, /rl/, /tr/, /nr/ and /rn/) in earlier IrE than it is today. It would appear, therefore, that schwa-epenthesis was more widespread and occurred in a greater range of phonetic environments, and that the present-day occurrence of liquid–nasal clusters (/lm/, /rm/) in syllable-final position is actually a more recent development.

In sum, the linguistic analysis of a corpus like CORIECOR allows both academic and wider audiences to delve into the linguistic heritage of different Irish regions from the eighteenth to the twentieth centuries. It can also be revealing in terms of tracking social and attitudinal changes across time, as we demonstrate in Sect. 5.

4 Digitizing Experiences of Migration

This section focuses on the DEM project, exploring the technological and other issues which connecting different data sets like DIPPAM and CORIECOR generate.

DEM, as already noted, was an interdisciplinary network project. It eventually included 20 partner scholars at third level institutions in Europe and the United States with expertise in digital humanities, historical studies, library, archive and information studies and linguistics.¹⁵ All the participants work with emigrant letters in various ways and key goals of the project included addressing the challenges surrounding the digitization and annotation of correspondence corpora more broadly with a view to: (a) building capacity in the use of corpus tools, and (b) initiating

¹⁵ Full details of the research network partners, together with the letter collections they are involved with, can be found on the project blog: <http://www.lettersofmigration.blogspot.co.uk>.

the process of interconnecting correspondence resources to encourage cross-disciplinary research and the reuse of the databases for wider audiences. Central to these goals is the development of a system of mark-up that conforms to the guidelines of the Text Encoding Initiative (TEI).¹⁶ This sort of annotation encodes various contextual, discursive, linguistic, physical and structural properties of the letters, thus enabling different aspects of the material to be explored. As well as allowing academic users to make more sophisticated searches, this also means that many different features of the collection can be presented through the application of visualization techniques that can encourage the interest of the general public as well as the creative and heritage industries (particularly film and the performing arts/museums).

4.1 Best Practices in the Building of Migration Databases

Although original migrant letters are not generally very accessible to researchers—never mind the general public—except by digital means, they form the basis of research in many disciplines, as previously noted. They also have great appeal outside the academy, to individuals interested in local or family history, or in learning about important historical events from a range of divergent perspectives (especially those of the *intrahistoria*). Thus, having access to a database like CORIECOR, which includes transcribed versions of letters such as those seen in Fig. 2.2, would allow members of the public to engage with contemporaneous accounts of the Great Irish Famine, for instance, which was precipitated by successive potato blights of the kind referred to here by Lewis Reford, the writer of the letter on the left.

While many collections have now been digitized in initiatives such as DIPPAM, not all are properly archived, some are reduplicated and others are in danger of being lost.¹⁷ The documentation and preservation of such letters is thus a particularly pressing need.

¹⁶ Mark-up here is used to describe a system for annotating a document which incorporates additional information in a manner that makes it syntactically distinguishable from the text itself.

¹⁷ Since the original CORIECOR data set was tied to DIPPAM until the former expanded its geographical range and time-depth, there are, naturally, many duplicates between these two collections, for instance.

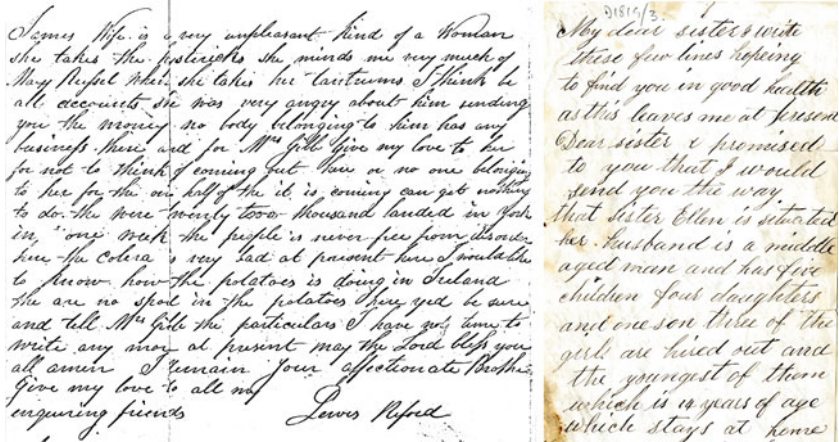


Fig. 2.2 Digital copies of: (a) PRONI T3028/B/5(2)—a letter written by Lewis Reford in Newburgh, Orange County, York State, to Fanny Reford, County Antrim on July 15th, 1849 (left); and (b) PRONI D1819/3—a letter written by Mary Quin, Barrytown, New York to her sisters in Stewartstown, Co. Tyrone on January 1st, 1873 (Reproduced by kind permission of the Deputy Keeper of Records, Public Record Office of Northern Ireland).

At present, most existing digital letter collections consist of unannotated versions of original manuscripts as in Fig. 2.2. The digitization process has naturally made letters such as these more accessible to academics and the general public, and it has also increased their searchability, at least to a certain extent. Unfortunately, however, emigrant correspondence projects have often evolved independently of one another, and although project teams have been successful in tackling important research questions relating to social history and immigration studies they have rarely joined forces, or engaged with stakeholder groups from other disciplines or indeed outside the academy. Moreover, relatively few projects have moved beyond the digitization stage to include even some basic mark-up in order to better exploit text content and enhance usability and searchability through the use of tools like *AntConc* described earlier in connection with CORIECOR. Different emigrant letter collections cannot easily interconnect if they are simply digitized without mark-up, and some search pathways through the material will remain unavailable if end-users are not encouraged to employ the right kinds of software tools to process this encoding.

It is within this research context that the DEM network set out to explore the digitization and annotation of historical emigrant letter collections, with a special focus, as CORIECOR also has, on letters of the Irish diaspora. Through a series of workshops, the scheme brought together scholars from different disciplines currently working with emigrant letters as a primary data source, to explore the digital potential of these iconic documents. In these workshops, participants examined how letters are being used across the disciplines, identifying where there are similarities and differences in transcription, digitization and annotation practices. The aims were to discuss issues and challenges surrounding digital corpus creation, to build capacity relating to correspondence mark-up and the use of corpus tools, and to initiate the process of inter-connecting resources to encourage cross-disciplinary research. Central to this goal was the development of a system of correspondence mark-up to represent the graphological and linguistic properties of the letters such as the encoding of misspellings (such as *hystericks* ('hysterics') in the Reford letter and *hopeing* ('hoping') in the Quin letter presented in Fig. 2.2) as well as those that reflect vernacular pronunciations like *Henery* ('Henry') and *worald* ('world'), already mentioned. It was also important to find principled mechanisms for marking up the contextual and physical properties of the letters themselves as objects. Such an annotation scheme would thus offer different layers of meaning and access points to the texts, allowing for more straightforward as well as more sophisticated searches. It would also permit the presentation of outputs from divergent correspondence corpora through meaningful visualizations with which wider audiences could engage.

4.2 Towards a System of Mark-up for Emigrant Letter Collections

The process of encoding is an intellectual activity, which involves thinking about which features of the document to represent, the relationship between those features and how they should be named, described and categorized in a formalized way. The way a text is encoded will reveal something about what we believe to be important or salient

about the original document. Encoding makes explicit our interpretation of a document and, as such, it is never a neutral process. The same text, then, can be encoded in many different ways, drawing out features that are most relevant to our own research interests, or projects. Encoding enriches the text, allowing us to look at it in new ways and from many disciplinary perspectives, offering diverse routes into the text and providing different layers of interpretation. For the DEM project, the encoding process was divided into the two stages elaborated on below, namely, document analysis (Sect. 4.2.1) and mark-up (Sect. 4.2.2).

4.2.1 Document Analysis

The first stage in our encoding project was to carry out a detailed analysis of the documents to be digitized and marked up, so as to identify features that a very wide range of end-users might consider to be important in and across the various emigrant letter collections. When analysing a selection of emigrant letters, the project partners were asked to consider two key questions:

- (a) What do different researchers use correspondence collections for?
- (b) What features of the letters are considered important across the disciplines?

In many ways, the document analysis stage of the process is the most challenging. Close examination of Fig. 2.3, for example, a letter by an Irish emigrant called Julia Lough, reveals a wealth of information that might be of interest to various kinds of users, for an equally wide range of purposes. A paper conservator or paper historian may, for instance, want to explore the quality of the paper or how the paper has been folded (vertically through the centre and horizontally one third from the top/bottom of the sheet). A graphologist, by contrast, may wish to capture information about the handwriting style, whether the document is the work of multiple hands, whether the original letter is written in pen or in pencil and how the document is structured. They may also want to

distinguish between writing that is contained in the margins and that which is found in the main body of the original document. By contrast, linguists will be more interested in capturing information about features of the letter's language, like those already mentioned as pertinent to the analysis of CORIECOR, such as spelling variations (for instance, *recived* ('received'), circled in hash marks in Fig. 2.3) and unusual syntax (*I dreamed you was*, circled in smaller hash marks two lines above the lower paper fold). They will likewise be keen to have a system for capturing the omissions or repetitions outlined in the discussion above of the John Stevenson Sinclair letter from CORIECOR (Sect. 2.1). Historians or non-academics interested in genealogy, on the other hand, might prefer to focus on any references to people, places and significant events such as those included in the texts of Figs. 2.2 and 2.3. For these and all end-users, in fact, there will also be contextual information that is not explicitly stated within the document content, but which is useful to capture nonetheless. This information may be inferred from the document itself

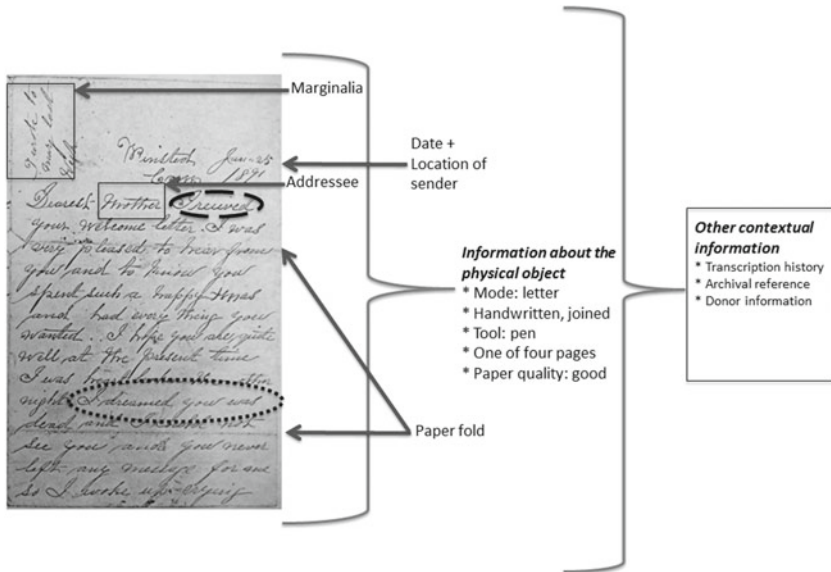


Fig. 2.3 Document analysis (letter by Julia Lough, Winsted, 1891 to her mother)

or it may be obtained from outside sources. We know, for instance, that the letter shown in Fig. 2.3 was written by someone called Julia Lough, a female Irish emigrant from a town called Meelick in what was then called Queen's County, Ireland, and is now County Laois. Additionally, previous research by immigration historians suggests that Julia had five sisters, three of whom, like Julia, emigrated to America in the 1870s and 1880s. Finally, we have metadata¹⁸ about where the Lough collection originated which it is also important to preserve, namely, that some of the letters were donated to Arnold Schrier, Professor Emeritus at the University of Cincinnati, by Canice and Eilish O'Mahony of Dundalk, Co. Louth, and the remaining letters were donated to Professor Kerby Miller, University of Missouri, by Edward Dunne and Mrs Kate Tynan of Portlaoise, County Laois.

Document analysis, then, is the starting point of the encoding process, looking at how the various disciplines, with their diverse range of potential research questions, use emigrant letters and how this process is also an important step to improving their utility beyond the academy (for family genealogy, for example, which one might readily see is very applicable to the Lough collection). Through document analysis, it becomes possible to identify where there were commonalities across the stakeholder types and where there were differences. Most importantly, of course, it was key to understanding where reduplication amongst the network partners operating on the same original or digital collections was taking place. There is not enough space in this chapter to go through all of the features we investigated, so just three features of the letters that were deemed to be important will be focused on here, namely, person, location and date. Having examined a selection of emigrant letters, taken from different archives, the network partners agreed that—where possible—the following information should be captured in relation to these three aspects of metadata (see Table 2.1). There are two things to note here: (a) the lists contained within Table 2.1 are not exhaustive (depending on the stakeholder, other or additional information may need to be captured) and (b) it is very rare that all of this information is available for any one

¹⁸Metadata is understood here to mean structured information that describes, explains and locates digital data.

Table 2.1 Person, location, date

Person (sender/addressee)	Location (sender/addressee)	Date
First name	Street	Day
Surname	Town	Month
Maiden name	Region/county	Year
Nickname(s)	Country	
Date of birth	Geographical coordinates	
Date of death	Additional information	
Sex		
Occupation(s)		
Social status		
Education		
Faith		
Relationships		
Places of residence		
Date of emigration + from/to		
Additional information		

letter. Indeed, it is more likely that we find ourselves working with a letter extract such as those in Figs. 2.2 and 2.3 rather than the entire document and these of course contain very little demographic information about the sender and/or the addressee.

4.2.2 Mark-up

Once the research network had examined the various ways in which letter collections are put to use, specifically identifying those features of letters that are important to a wide range of end-users, the next stage was to agree on how to model that information using TEI mark-up. This step in the process thus formalized and standardized this metadata and allowed the letter collections to potentially interconnect.

The research network used the TEI P5 Guidelines to carry out the encoding process in the same way that Mearns et al. (this volume) have dealt with their collection of spoken data in transcribed form.¹⁹ These rather different projects were both able to make use of the TEI system of mark-up precisely because the guidelines are designed to work with

¹⁹ For full details of the TEI P5 Guidelines, see <http://www.tei-c.org/Guidelines/P5/index.xml>.

different kinds of digitized texts, across all disciplines of the humanities. Specifically, the guidelines:

make recommendations about suitable ways of representing those features of textual resources which need to be identified explicitly in order to facilitate processing by computer programs. In particular, they specify a set of markers (or tags) which may be inserted in the electronic representation of the text, in order to mark the text structure and other features of interest (TEI Consortium, p. xxiii).

The TEI Special Interest Group on Correspondence (established in 2008)

seeks to bring together scholars interested in creating digital scholarly editions of correspondence [...] to discuss and develop sample tagsets (including suggesting additions/modifications to the TEI Guidelines) for varying forms of correspondence (TEI SIG: Correspondence wiki).

Amongst other things, their work has involved developing special purpose elements for correspondence-specific metadata, which are organized under the element `<correspDesc>` (see Fig. 2.4), allowing features of letters

```
<profileDesc>
  <ct:correspDesc>
    <ct:sender>
      <persName key="LOUGHPers_0001"/>
    </ct:sender>
    <ct:addressee>
      <persName key="LOUGHPers_0002"/>
    </ct:addressee>
    <ct:placeSender>
      <placeName key="LOUGHPlace_0001"/>
    </ct:placeSender>
    <ct:dateSender>
      <date when="1891-10-25"/>
    </ct:dateSender>
  </ct:correspDesc>
</profileDesc>
```

Fig. 2.4 Example mark-up (person, location and date)

such as sender and addressee to be represented in a standardized way. At present, it is recommended that these special purpose elements are embedded under `<profileDesc>` within the TEI header. The purpose of the `<profileDesc>` section of the TEI header is to provide a text profile

containing classificatory and contextual information about the text, such as its subject matter, the situation in which it was produced, the individuals described by or participating in producing it, and so forth (TEI Consortium, p.17)

as can also be seen in Fig. 2.4.²⁰

The mark-up shown in Fig. 2.4 captures information about the sender `<ct:sender>`, the addressee `<ct:addressee>`, and the date of the letter `<ct:dateSender>`. Embedded within `<ct:sender>` is the element `<persName>` which contains a unique identifier that corresponds to an element within a separate XML personography file (in this case LOUGHPers_0001). This individual personography file contains more detailed information about the sender (such as that outlined in Table 2.1). Similarly, embedded within `<ct:placeSender>` is the element `<placeName>` which contains a unique identifier that corresponds to an element within a separate XML placeography file (in this case LOUGHPlace_0001). This individual placeography file contains more detailed information about the sender's location (such as that also described in Table 2.1). Information about the addressee `<ct:addressee>` is then organized in the same way.²¹

The personography and placeography files are effectively the same as 'authority files', a term used by archivists and librarians to describe bibliographic master files. There is one personography file (or authority file) for each person and there is one placeography file (or authority file) for each place. These authority files for person and place are each given a

²⁰It should be noted that the mark-up being proposed in this chapter is what was discussed and agreed on during the workshops. Since writing this chapter, the TEI `<correspDesc>` proposal has been finalized and now centres around `<correspAction>` and `<correspContext>` elements inside `<correspDesc>`, for describing details about the sending and reception of a letter as well as the context in which the letter occurs. For details about using TEI to model correspondence, see the Correspondence SIG wiki (<http://wiki.tei-c.org/index.php/SIG:Correspondence>). For XML files showing applications of the `<correspDesc>` proposal, go to <https://github.com/TEI-Correspondence-SIG/correspDesc>.

²¹A 'personography file' simply means the annotation and preservation of biographical data and a 'placeography file' is very similar but refers to data referring to locations rather than people.

unique identifier, which is then referred to in the mark-up. In other words, LOUGHPers_0001, for instance, corresponds to an element within a separate file which contains information about this particular participant (Julia Lough) such as date of birth, first name, surname, maiden name, nicknames (including all spelling variations), sex, occupations, date of emigration, date of death and so on. Similarly, LOUGHPlace_0001 corresponds to an element within a separate file that includes information about this particular place (Winsted) incorporating, for instance, its geographical coordinates. Having separate personography and placeography files for each person and place mentioned in the original documents and allied materials makes it much easier to manage changes to related metadata at a later date since it is clearly easier to change one master document—that is, the personography or placeography file—than it is to change hundreds of individual documents.

In addition to the personography and placeography information, <ct:dateSender> contains details of when the letter was dated. There are different ways to capture this information within the header.²² If the letter contains a date, then the day, month and year can be represented in the mark-up (as in Fig. 2.4). However, quite often an exact date is missing and it is up to the reader to make an educated guess as to when the letter was written. In such instances, the <notBefore> and <notAfter> attributes can be used to place the letter within an approximate time frame, for example <date notBefore="1800" notAfter="1899"/>.

It is not possible, in the space of this chapter, to discuss how the personography and placeography information can be modelled using TEI. However, further information about modelling emigrant letter texts, together with links to example XML files, can be found on the project blog.²³

4.3 Interconnecting Resources

Having agreed on how best to model information relating to person, location and date, within the TEI header, it was then possible to start the process of interconnecting some of the emigrant letter collections that the project partners are involved with in a similar manner to that of

²²A TEI header supplies the descriptive and declarative information comprising an electronic title page for every document which is considered to be TEI-conformant.

²³See <http://www.lettersofmigration.blogspot.co.uk>.

(a)	#	id	url	date	sender	place_of_sender	addressee	place_of_addressee
	1	+472436	http://umedia.lib.umn.edu/node/472436	April 30, 1924	Grebenstchikoff, George	New York City, New York	Nicholas R	
	2	+710673	http://umedia.lib.umn.edu/node/710673	June 6, 1950	Budzka, Eduard	Wentorf, Germany	Klara and Vaclav Pa	
	3	+88790	http://umedia.lib.umn.edu/node/88790	September 10, 1957	Paikens, Helena	Minneapolis, Minnesota	her mother-	
	4	+88811	http://umedia.lib.umn.edu/node/88811	January 19, 1929	Neprytsky-Hranovsky, Serhii	Berezhtsi, Ukraine		
	5	+88833	http://umedia.lib.umn.edu/node/88833	January 12, 1912	Aalto, Bert	Big Falls, Minnesota	his friend	
	6	+88831	http://umedia.lib.umn.edu/node/88831	August 26, 1911	Aalto, Bert	Big Falls, Minnesota	his friend Hilma A	
	7	+472442	http://umedia.lib.umn.edu/node/472442	May 16, 1936	Roerich, Nicholas	Kullu, India	George and Tatiana	
	8	+88793	http://umedia.lib.umn.edu/node/88793	December 11, 1898	Fazio, Lucia	Hoboken, New Jersey	Alessando f	
	9	+553855	http://umedia.lib.umn.edu/node/553855	July 5, 1927	Kovač, T. Paul	Hawleyville, Connecticut	his mother	
	10	+472455	http://umedia.lib.umn.edu/node/472455	November 8, 1949	Douline, Nikolay	Southbury, Connecticut	Ge	
	11	+710636	http://umedia.lib.umn.edu/node/472455	August 14, 1914	Arthur Sch.	probably Buenos Aires	Emilie Wehle	Vi
	12	+552682	http://umedia.lib.umn.edu/node/552682	1948-09-30	Petris, Antonietta	Montreal, Canada	Loris Palm	
	13							
(b)	40	+	<ct:correspDesc xmlns:ct="http://wiki.tei-c.org/index.php/SIG:Correspondence/task-force-correspDesc" n="710673">					
	41	+	corresp="http://umedia.lib.umn.edu/node/710673">					
	42	+	<ct:sender>Budzka, Eduard</ct:sender>					
	43	+	<ct:addresssee>Klara and Vaclav Panucevich</ct:addresssee>					
	44	+	<ct:placeSender>Wentorf, Germany</ct:placeSender>					
	45	+	<ct:placeAddressee>Chicago, Illinois</ct:placeAddressee>					
	46	+	<ct:dateSender when="1950-06-06"/>					
	47	+	</ct:correspDesc>					
	48	+	<ct:correspDesc xmlns:ct="http://wiki.tei-c.org/index.php/SIG:Correspondence/task-force-correspDesc" n="88790">					
	49	+	corresp="http://umedia.lib.umn.edu/node/88790">					
	50	+	<ct:sender>Paikens, Helena</ct:sender>					
	51	+	<ct:addresssee>her mother-in-law, Anna Paikens</ct:addresssee>					
	52	+	<ct:placeSender>Minneapolis, Minnesota</ct:placeSender>					
	53	+	<ct:placeAddressee>Iencini, Latvia</ct:placeAddressee>					
	54	+	<ct:dateSender when="1957-09-10"/>					
	55	+	</ct:correspDesc>					
	56	+						
	57	+						

Fig. 2.5 (a) An example of the raw metadata from the letter collection held at the IHRC, University of Minnesota; (b) TEI compliant metadata

the ENROLLER scheme at the University of Glasgow, which created an integrated online repository of electronic resources for the study of language and literature.²⁴ We focused on two collections, in particular: (a) the IED subset of DIPPAM and (b) letters from the Digitizing Immigrant Letters (DIL) project at the Immigration History Research Centre at the University of Minnesota—a collection of about a hundred letters by migrants and their families in Europe and North America.²⁵ The first step was to standardize the metadata that is available for these two collections. An example of the raw information of this kind can be seen in Fig. 2.5a. The standardized metadata was then passed to Peter Stadler (Universität Paderborn) to be made TEI-compliant, as represented in Fig. 2.5b. This involved transforming generic TEI to the more specific <correspDesc> compatible TEI variety via XSLT.²⁶

²⁴ See <https://enroller.nesc.gla.ac.uk>.

²⁵ See <http://www.ihrc.umn.edu/research/dil/aboutDIL.htm>.

²⁶ In other words, Extensible Stylesheet Language Transformations, which is a language for transforming XML documents into other XML documents, or other formats such as HTML for web pages or even plain text.

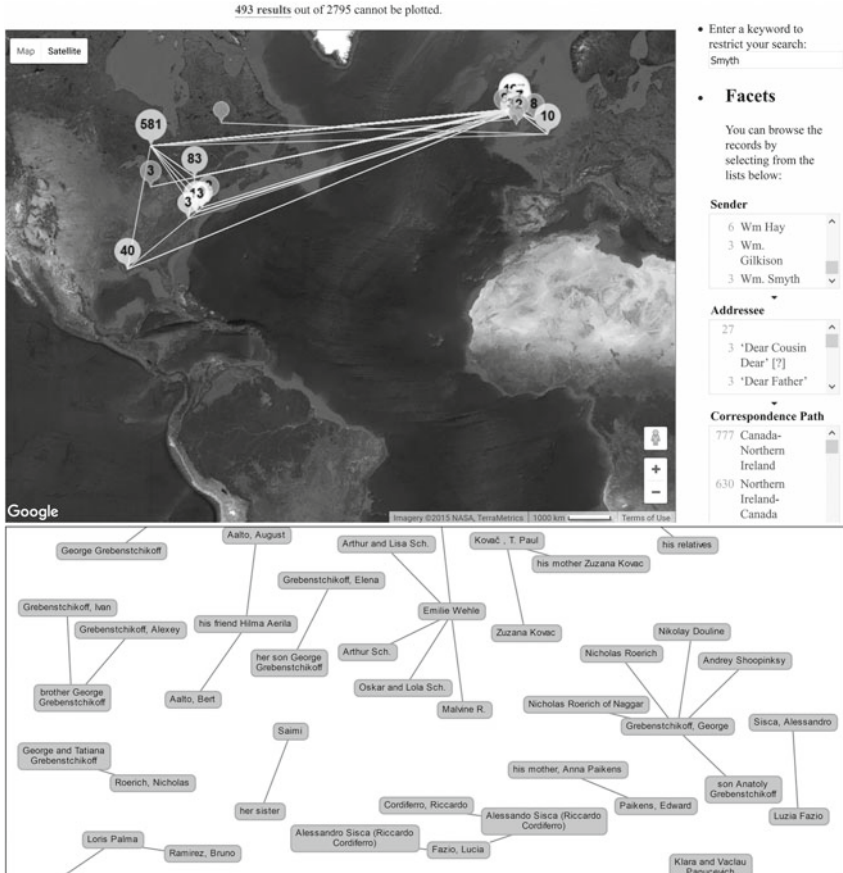


Fig. 2.6 Visualization 1 (top): The location and movement of migrants. Visualization 2 (bottom): Letter writing networks

Through standardizing this metadata and making it TEI-compliant, it was then possible for Niall O’Leary (a freelance programmer) to create a range of visualizations, exploring aspects of migration that give insights into migrant relations and mobility patterns, such as the movement of migrants over time or the establishment of letter-writing networks (see Fig. 2.6). These can readily be exploited for public engagement

purposes and indeed have already been used successfully in the Herbert exhibition described in Sect. 5.1.²⁷

What this encoding process highlighted, in particular, is the potential for working with header information (in this case information embedded within <correspDesc>) relating to person (sender and addressee), location and date, without necessarily having access to the letter itself. One of the biggest challenges that came out of the network process related to accessibility of letter collections and issues to do with intellectual property. In other words, it is often difficult to get access to collections and even more difficult to make collections freely available online to either the public or the academy, especially when working across disciplines and across cultures and the different legal frameworks of diverse national jurisdictions. By focusing on metadata about the letter (rather than the letter itself) there were fewer barriers to overcome with regard to interconnecting resources. This is, of course, just the first stage in terms of developing fully interoperable correspondence resources, but hopefully—with further funding—more letter collections will be able to interconnect in this manner using the model developed during this project.

5 Irish Migration Databases as Impact Tools

Thus far we have emphasized the importance to the aims of both the CORIECOR and DEM projects of creating databases that impact upon researchers operating within different disciplines. We have also highlighted the fact that a key objective is to design correspondence corpora that will engage the general public. This section reports on three initiatives which trialled these databases in England and Northern Ireland for exactly this purpose: (a) The *Leaving, Crossing, Arriving* exhibition in May 2014 at the Herbert Art Gallery and Museum, Coventry (Sect. 5.1); (b) a public lecture at the MCMS in October 2014 (Sect. 5.2) and (c) the *From Home to Here* exhibition and

²⁷The visualizations created for this project are particularly useful for engaging a wider audience. For full details and further examples, see the ‘*Leaving, Crossing, Arriving*’ page on Niall O’Leary’s site: <http://development.nialloleary.ie/correspondence/correspondence.php>.

accompanying booklet (Sect. 5.3) relating to another public event at the MCMS in June 2015 which has just gone on a tour of public libraries across Northern Ireland.²⁸

5.1 The *Leaving, Crossing, Arriving* Exhibition

The Herbert exhibition promoted access to information within and across digitized emigrant letter collections and was specifically designed for non-academic audiences. To do this, the exhibition exploited some of the visualizations mentioned in the previous section arising from the DEM project, demonstrating aspects of migration such as letter-writing networks, the movement of migrants over time and how often migrant families like the Quins, the Refords and the Loughs, whose correspondence has already been referred to, corresponded with one another. The visualizations helped the public to contextualize the individual letters that were part of the exhibition within the larger context of nineteenth-century migration (see Fig. 2.7).

The Herbert exhibition was also used as a stimulus for creative work by local primary- and secondary-school students, encouraging them to examine language change and identity as aspects of migratory processes as well as addressing issues of integration and assimilation which relate to subject areas within the National Curriculum for England.²⁹ Many of the children who took part came from immigrant families themselves. As such, the letters—and the exhibition itself—did not just serve an educational purpose, but also provided an opportunity for the children to reflect upon the degree to which their own experiences were similar to or different from those reported in the historical correspondence. The Julia Lough letter, for example, was given to students to work with in class before attending the exhibition. In the letter, an extract of which is given

²⁸The lecture was delivered by Karen Corrigan and Adam Mearns under the auspices of the MBDNL project already mentioned. It was part of a day-long event they led on this theme at the *Fifteenth Literature of Irish Exile Autumn School* at the MCMS, a key institutional partner in the CORIECOR, DEM and MBDNL initiatives. For details of the event see http://www.qub.ac.uk/cms/events/15th_LIE_2014/LIE_Oct_2014.htm. The lecture and details of the MCMS booklet can be found at the MBDNL website: <http://research.ncl.ac.uk/ni-language-migration>.

²⁹See <https://www.gov.uk/government/collections/national-curriculum>.

Leaving, Crossing, Arriving

Digitising Experiences of Migration: the development of interconnected letter collections

A Research Networking Project funded by the Arts and Humanities Research Council



Why this project?

Although many emigrant letter collections have now been digitised, not all are properly archived; some are reduplicated and others are in danger of being lost. The documentation and preservation of such letters is a particularly pressing need.

Emigrant correspondence projects have almost always evolved independently

Fig. 2.7 Sample of foamex board from the Herbert exhibition

in (5), Julia talks about a dream she had in which her mother was dead. She expresses feelings of homesickness and anxiety about loved ones back in Ireland. It is through this and similar letters that Julia helps to maintain familial bonds over time and distance.

- (5) I hope you are quite
Well at the present time
I was heart broken the other
night I dreamed you was
dead and I could not
See you and you never
left any message for me
so I woke up crying
[new page]
and I was so frightened till

I realised it was only a
dream. I hope I Shall
meet you once more in life
and have a happy time
again. let me know are [...]

[Extract of letter by Julia Lough, Winsted, 1891 to her mother in
Meelick, Ireland]

The students were given a range of tasks relating to various letter extracts, which included looking at the language of emotions, as well as aspects of style, register and genre, and, finally (depending on the age, background and needs of the students) they were also encouraged to talk about their own experiences of emigration and to produce letters themselves which reflected this.³⁰

5.2 The Mellon Centre for Migration Studies Public Lecture

The MCMS public lecture shared similar goals to those which motivated the Herbert exhibition regarding the comparison of migratory experiences as the title suggests: ‘What Do “Young Irelanders” and “New Kids on the Block” have in Common?’ Its particular focus was on comparing the narratives of historical migrants from Ireland (‘Young Irelanders’) with those of recent migrants to Northern Ireland (‘New Kids on the Block’), whose arrival can be linked to the dividends of the 1990s Peace Process, as well as to the expansion of the European Union from 2004 onwards.³¹ The historical data set used in this public engagement event derives from CORIECOR. This resource has been united in the AHRC-funded MBDNL project referred to above with a new corpus of sociolinguistic

³⁰ Full details of the Herbert exhibition, together with images of the display boards constructed for it, can be found on the project blog at <http://lettersofmigration.blogspot.fr/p/symposium-and-exhibition.html>.

³¹ The ‘Young Irelanders’, which included Thomas Davis, its chief organizer, was a sociocultural and political movement in nineteenth-century Ireland leading to changes in Irish nationalism and leading to an abortive rebellion in 1848. Many of its leaders were found guilty of sedition and sentenced to penal transportation to Van Dieman’s Land.

interviews conducted between 2013 and 2014 which record primary and post-primary pupils living in three different regions of Northern Ireland.³² The interview protocol was constructed to elicit ethnographic information as well as spoken data for linguistic analysis of the kind detailed in Sects. 2.1 and 3.1. In addition, it was designed to provide insights into the formation of identity amongst migrants in language contact settings and the extent to which the identities they project may or may not be related to their degrees of integration in different types of community within Northern Ireland. In that sense, therefore, it compares very well with the letters in CORIECOR and indeed in DEM, which can also be used to test degrees of alienation in migrant contexts, such as that reported in the extract in (5) from the Julia Lough letter. The public lecture thus compared the CORIECOR text in (6), which was written in 1840 in Virginia, USA, by Moses Paul, with the extract in (7), from a 2013 interview with an 18-year-old Polish migrant ‘Natazsa Pawelski’.

(6) I am now nearly two years in this republic, and have not recd. the first scrape of your pen—not even on the back of a newspaper. Why is this? am I beneath your notice because I am turned American [...] From what I have often heard sister saying, I presume she is still as prejudiced as ever against America & Americans [...] this is wrong—persons living in Europe cannot have any idea of this country & should not condemn it so unmercifully without judge or jury—certainly we differ from you in some points very materially, but it is not worth talking of.

[Extract from CORIECOR’s letter by Moses Paul, Virginia, dated 1840]

(7) I don’t regret ehm coming here ‘cause I’ve met some lovely people and I don’t think I could go back and live in Poland again; it’s so different, even the people [...] I’m not saying that Polish people aren’t nice, but ehm like compared to here, they’re very intolerant.

[Extract from the MBDNL corpus interview with ‘Natazsa Pawelski’, Armagh, dated 2013]

³² See <http://research.ncl.ac.uk/ni-language-migration>.

In each case, the subtexts indicate that migrants have reached high levels of integration within the host community and, indeed, that they have begun to develop negative views about the land from which they originated. Thus, Moses Paul in (6) describes himself as having ‘turned American’ while ‘Natazsa’ notes her reluctance to return to Poland in (7).

Interestingly, some correspondents and interviewees are more like Julia Lough in their orientation. For example, the extract in (8) is from an early twentieth-century letter in which Andrew Dunn, a migrant to Ontario, indicates that if he had the chance he ‘would go back [to Ireland] again’. His perspective chimes very well with that of a family member of another Polish teenager living in Armagh, ‘Polly Wiczorek’, who describes in (9) the views held by one of her siblings who also ‘wants to go back’.

- (8) all the same if I could get a little Farm I would go back again, this place is alright so long as everything goes well but when one gets sick it is not so good.

[Extract from CORIECOR’s letter by Andrew Dunn, Ontario, dated 1929]

- (9) she was depressed ’cause she had all her friends back in Poland. She was twelve [...] Yeah so she didn’t like it, but she does she doesn’t like it still she wants to go back.

[Extract from the MBDNL corpus interview with ‘Polly Wiczorek’, Armagh, dated 2014]

5.3 The *From Home to Here* Exhibition and Booklet

The MBDNL project team (Karen Corrigan, Adam Mearns and Jennifer Thorburn) ran a free public exhibition called *From Home to Here: Stories of Migration Old and New* at the MCMS Library in June 2015. It was dedicated to addressing issues of identity, language and migration relating to the North of Ireland and was presented from both synchronic and diachronic perspectives. The exhibition included a display of objects that both contemporary immigrants and historical emigrants might have brought with them to represent the cultures, communities and languages they had left behind. The curation, selection and interpretation

of these artefacts were done by local primary and post-primary school children and were facilitated by our outreach partnerships with different educational institutions across the region begun in 2008 during the fieldwork for Corrigan (2010). The exhibition had both physical and virtual elements, with the latter being achieved by devoting dedicated space on the MBDNL's project website to materials that did not readily lend themselves to 3D display. These included audio clips, documentaries, lectures and news items which could all be interacted with in the MCMS Library site at computer booths specially kitted out for this purpose.³³ *From Home to Here* also incorporated informational panels organized into four different themes ('journey', 'prosperity', 'culture and heritage' and 'home and belonging') generated by an *AntiConc* review of text topics in the corpora described in the previous section. Digital images of the CORIECOR letters, such as those in Fig. 2.2, as well as political cartoons, news reports and memorials, were then used to illustrate the historical extracts. Conversations on these topics from the sociolinguistic interviews were illustrated with contemporary photos representing migrant languages and cultures in what McDermott (2011) terms 'the public space' (for example, arts projects and community festivals, such as the Belfast *mela*).³⁴

The exhibition was accompanied by a booklet with more text extracts and images than the limited physical space of the MCMS Library could accommodate. It is prefaced with a narrative description of migration involving this region from the earliest times to the present day which is copiously illustrated and written in an accessible style. It also contains eye-catching titbits of information as well as a nuanced presentation of historical fact so as to pique the interest of wider publics (Fig. 2.8).

Threaded through the narrative are discussions of the key economic and sociocultural push–pull factors which precipitate population movements more generally as well as an outline of their linguistic consequences both for the migrant and for the community which they leave

³³ These items included BBC news reports, public lectures and national archive materials and the like that were in the public domain or for which permissions could be obtained. The virtual exhibition in its entirety can still be viewed at the *From Home to Here* exhibition website: <https://irishmigration.wordpress.com>.

³⁴ For information on the Belfast *mela* festival, see <http://2015.belfastmela.org.uk>.



Fig. 2.8 Extracts from the *From Home to Here* booklet

behind. A good case in point here is the report which features in the booklet (Fig. 2.9) by Séan Ó Dúbhda, a respondent to the Irish Folklore Commission's 1955 Emigration Questionnaire, which names 'the American Letter' as a key force in the shift from Irish to English in Ireland during the nineteenth century (see Corrigan 1992).

The booklet is accompanied by a CD which not only contains audio files of the conversations on each of the exhibition's themes from participants in the sociolinguistic interviews, but also includes dramatic readings of the CORIECOR letters used in the text. These were produced by local drama students who are from the dialect area in Ulster where the letter writer was also originally from and who therefore have accents that the wider public will recognize as typical of those regions.

Taken together, the exhibition, booklet and website pages relating to the *From Home to Here* initiative serve to demonstrate the disparities and synergies between contemporary and historical processes of migration into and out of the North of Ireland. Moreover, they do so with a keen eye on techniques—like the dramatic reading of the CORIECOR letters or the juxtapositioning of texts on education with quirky images (Fig. 2.10)—which will engage non-academic audiences to connect with these databases in ways that go beyond simply supplying them with a digital image or transcription of the original. What is more, once the mark-

279

Heading 7.

I think the American letter helped to anglicise this country to a great extent and gave the people a greater desire to learn English and to keep the children at school so as to learn it and have some knowledge of that language.

I often heard my father to say that when he was a young lad rising up nearly every letter that came from America at that time urged and exhorted the parents to try and teach English to the children.

'S gcúntais Dé muintir Béarla dos na leanbhais, is ná bídis dall ar nós na n-asal a' teacht anso amaic.

That was some of the talk in the letters.

Every youngster was a potential emigrant.

There were many cases where parents who could not speak English gave the rod to their children across the shinbones near the fire because they spoke Irish.

The American letter was the most cause of it.

Was it because of patriotism or a geographical accident the Irish language survived?

Fig. 2.9 Extract from Séan Ó Dúbhda's response to the 1955 Irish Folklore Commission's Questionnaire (Reproduced by kind permission of the National Folklore Collection, University College Dublin).



Fig. 2.10 Extract 3 from the *From Home to Here* booklet

ing up of CORIECOR as TEI-compliant texts using the best practices devised in the DEM project is complete, there should be no reason why CORIECOR could not be made available so as to include audio and images of the kinds described here that will really enhance the public's experience of using them.

The exhibitions, public lecture and booklet outlined in this section are best regarded as test cases for the use of migration databases in educational and heritage contexts. Nevertheless, they do represent the first successful attempts to do so, as far as we are aware, using corpora that relate to migratory processes connected with the island of Ireland. Moreover, it is hoped that they can serve as a model for future public engagement initiatives of this type.

6 Conclusion

This chapter examined the DIPPAM and CORIECOR projects with a view to highlighting their value as databases for undertaking linguistic analyses of correspondence corpora that are socially situated in the his-

torical contexts in which the letters were first composed. We also reviewed the key findings of the DEM project which established a set of best practices for the collation and annotation of these kinds of databases using world standards for encoding correspondence. Not only do these principles ensure the preservation of the data (and metadata) longer term as well as interoperability between the quite divergent digital letter corpora available nowadays, permitting end-use by various academic disciplines, they also make it possible for the data to be presented more attractively in primary/post-primary educational and heritage contexts. As already noted, TEI-conformant XML documents allow the databases to be searched not only for the kinds of contextual historical information provided by DIPPAM but also for an extensive range of features, including those that are linguistically relevant as well as others that might be of wider interest, such as graphological practices. Correspondence corpora also readily capture the migrant experience—their rationale for leaving, their attitudinal dispositions towards the new communities in which they find themselves as well as their views about the old social networks which their migration has interrupted. These sociological and artefactual aspects of the letters are key to the usefulness of emigrant letter corpora in engaging non-academic audiences in live events, displays and publicly oriented books.

It is hoped that the creation of correspondence corpora using the tools described in this chapter will mark the beginning of an era in which a diverse range of annotated correspondence corpora can be fully exploited by different types of end-user because they have become interoperable, user-friendly resources in the manner of the ENROLLER scheme which serves exactly this function for many of the corpora described in Beal et al. (2007a, b).

References

Books and Articles

- Amador-Moreno, Carolina P., and Kevin McCafferty. 2012. Linguistic identity and the study of Irish letters: Irish English in the making. *Lengua y Migración* 4(2): 25–42.

- . 2016. “Sure this is a great country for drink and rowing at elections”: pragmatic markers in the *Corpus of Irish English Correspondence, 1750–1940*. In *Pragmatic Markers in Irish English*, eds. Carolina P. Amador-Moreno, Elaine Vaughan, and Kevin McCafferty, 270–291. Amsterdam: John Benjamins.
- . 2015. “[B]ut *sure* its only a penny after all”: Irish English discourse marker *sure*. In *Transatlantic Perspectives on Late Modern English*, ed. Marina Dossena, 179–197. Amsterdam: John Benjamins.
- Beal, Joan C., Karen P. Corrigan and Hermann L. Moisl (eds). 2007a. *Creating and Digitizing Language Corpora: Vol.1, Synchronic Databases*. Basingstoke: Palgrave Macmillan.
- (eds). 2007b. *Creating and Digitizing Language Corpora: Vol.2, Diachronic Databases*. Basingstoke: Palgrave Macmillan.
- Biber, Douglas. 2003. Variation among university spoken and written registers: a new multi-dimensional analysis. In *Corpus Analysis. Language Structure and Language Use*, eds. Pepi Leisty, and Charles F. Meyer, 47–70. Amsterdam: Rodopi.
- Biber, Douglas, and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.
- Close, Joanne and Bas Aarts. 2010. Current change in the modal system of English: a case study of *must*, *have to* and *have got to*. In *The History of English Verbal and Nominal Constructions*. Volume 1 of *English Historical Linguistics 2008. Selected Papers from the 15th International Conference on English Historical Linguistics (ICEHL 15), Munich, 24–30 August 2008*, eds., Ursula Lenker, Judith Huber and Robert Mailhammer, 165–181. Amsterdam: John Benjamins.
- Corrigan, Karen P. 1992. “I gcuntas Dé múin Béarla do na leanbháin”: eisimirce agus an Ghaeilge sa naoú aois déag (“For God’s sake teach the children English”: emigration and the Irish language in the nineteenth century). In *The Irish World Wide: History, Heritage, Identity. Vol. 2, The Irish in the New Communities*, ed., Patrick O’Sullivan, 129–142. Leicester: Leicester University Press.
- . 2003. The ideology of nationalism and its impact on accounts of language shift in nineteenth century Ireland. *Arbeiten aus Anglistik und Amerikanistik* 28(2): 201–230.
- . 2010. *Irish English, Volume 1: Northern Ireland*. Edinburgh: Edinburgh University Press.
- Daly, Mary E. 1990. Literacy and language change in the late nineteenth and early twentieth centuries. In *The Origins of Popular Literacy in Ireland: Language Change and Educational Development 1700–1920*, eds. Mary E. Daly, and David Dickson, 153–166. Dublin: Anna Livia Ltd.

- de Rijke, Persijn M. Forthcoming. “I intend to try some other part of the world”: evidence of schwa-epenthesis in the historical letters of Irish emigrants. In *Voice and Discourse in the Irish Context*, eds. Diana Villanueva Romero, Carolina Amador-Moreno, and Manuel Sánchez García. Basingstoke: Palgrave Macmillan.
- Devlin Trew, Johanne. 2013. *Leaving the North: Migration and Memory, Northern Ireland, 1921–2011*. Liverpool: Liverpool University Press.
- Dossena, Marina. 2012. The study of correspondence: theoretical and methodological issues. In *Letter Writing in Late Modern Europe*, eds. Marina Dossena, and Gabriella del Lungo Camiciotti, 13–30. Amsterdam: John Benjamins.
- Earle, Peter G. 1964. Unamuno and the theme of history. *Hispanic Review* 32(4): 319–339.
- Elliott, Bruce S., David A. Gerber, and Suzanne M. Sinke. 2006. Introduction. In *Letters across Borders: The Epistolary Practices of International Migrants*, eds. Bruce S. Elliott, David A. Gerber, and Suzanne M. Sinke, 10–25. New York: Palgrave Macmillan.
- Elsaß, Stephan. 2005. *Sprachgeschichte von unten. Untersuchungen zum geschriebenen Alltagsdeutsch im 19. Jahrhundert*. Berlin: Walter de Gruyter.
- Erickson, Charlotte. 1972. *Invisible Immigrants: The Adaptation of English and Scottish Immigrants in Nineteenth-Century America*. London: Weidenfeld & Nicolson, The London School of Economics and Political Science.
- Fehringer, Carol, and Karen P. Corrigan. 2015. “You’ve got to sort of eh hoy the Geordie out”: Modals of obligation and necessity in 50 years of Tyneside English. *English Language and Linguistics* 19(2): 355–381.
- Fitzmaurice, Susan 2007. Questions of standardization and representativeness in the development of social networks-based corpora: the story of the network of eighteenth-century English texts. In *Creating and Digitizing Unconventional Corpora. Volume 2: Diachronic Corpora*, eds., Joan C. Beal, Karen P. Corrigan and Hermann L. Moisl, 49–81. Basingstoke: Palgrave Macmillan.
- Fitzpatrick, David. 1994. *Oceans of Consolation: Personal Accounts of Irish Migration to Australia*. Cork: Cork University Press.
- Fitzgerald, Patrick, and Brian Lambkin. 2008. *Migration in Irish History, 1607–2007*. London: Palgrave Macmillan.
- Hendrickx, Iris, Michel Génèreux, and Rita Marquilha. 2011. Automatic pragmatic text segmentation of historical letters. In *Language Technology for Cultural Heritage*, eds. Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, 135–152. Berlin and Heidelberg: Springer.

- Kamphoefner, Walter D., Wolfgang Helbich, and Ulrike Sommer. 1988. *News from the Land of Freedom: German Immigrants Write Home*. Ithaca and London: Cornell University Press.
- Kennedy, Liam, Paul S. Ell, E.M. Crawford, and L.A. Clarkson, eds. 1999. *Mapping the Great Irish Famine*. Dublin: Four Courts Press.
- Knight, Dawn. 2011. *Multi-Modal Corpora*. Birmingham: University of Birmingham.
- Krug, Manfred. 2000. *Emerging English Modals: A Corpus-based Study of Grammaticalisation*. Berlin and New York: Mouton de Gruyter.
- Leech, Geoffrey. 2003. Modality on the move: the English modal auxiliaries 1961–1992. In *Modality in Contemporary English*, eds. Roberta Facchinetti, Manfred Krug, and Frank Palmer, 223–240. Berlin: Mouton de Gruyter.
- Lillis, Theresa. 2013. *The Sociolinguistics of Writing*. Edinburgh: Edinburgh University Press.
- Mair, Christian. 2006. *Twentieth-Century English: History, Variation and Standardization*. Cambridge: Cambridge University Press.
- McCafferty, Kevin. 2016. What can one short emigrant letter tell us about Irish English? *Ulster Folklife* 62.
- McCafferty, Kevin and Carolina P. Amador-Moreno. 2012. “I will be expecting a letter from you before this reaches you”: a corpus-based study of *will/shall* variation. In *Letter Writing in Late Modern Europe*, eds. Marina Dossena, and Gabriella del Lungo Camiciotti, 179–204. Amsterdam: John Benjamins.
- . 2014. [The Irish] find much difficulty in these auxiliaries [...], putting *will* for *shall* with the first person”: The decline of first-person *shall* in Ireland, 1760–1890. *English Language and Linguistics* 18(3): 407–429.
- McDermott, Philip. 2011. *Migrant Languages in the Public Space: A Case Study from Northern Ireland*. Berlin/London: LIT Verlag.
- Miller, Kerby A. 1985. *Emigrants and Exiles: Ireland and the Irish Exodus to North America*. Oxford: Oxford University Press.
- . 2008. *Ireland and Irish America. Culture, Class, and Transatlantic Migration*. Dublin: Field Day Files.
- Milroy, Lesley. 1987. *Language and Social Networks*, 2nd edn. Oxford: Blackwell.
- Montgomery, Michael B. 1995. The linguistic value of Ulster emigrant letters. *Ulster Folklife* 41: 26–41.
- . 2001. On the trail of early Ulster emigrant letters. In *Atlantic Crossroads: Historical Connections Between Scotland, Ulster and North America*, eds. Patrick Fitzgerald, and Steve Ickringill, 13–26. Newtownards: Colourpoint Books.

- Ó Gráda, Cormac. 2000. *Black '47 and Beyond: The Great Irish Famine in History, Economy, Memory*. Princeton: Princeton University Press.
- O'Sullivan, Patrick, ed. 1992a. *The Irish World Wide: History, Heritage, Identity: Vol. 1, Patterns of Migration*. Leicester: Leicester University Press.
- , ed. 1992b. *The Irish World Wide: History, Heritage, Identity: Vol. 2, The Irish in the New Communities*. Leicester: Leicester University Press.
- , ed. 1992c. *The Irish World Wide: History, Heritage, Identity: Vol. 3, The Creative Migrant*. Leicester: Leicester University Press.
- , ed. 1992d. *The Irish World Wide: History, Heritage, Identity: Vol. 4, Irish Women and Irish Migration*. Leicester: Leicester University Press.
- , ed. 1992e. *The Irish World Wide: History, Heritage, Identity: Vol. 5, Religion and Identity*. Leicester: Leicester University Press.
- , ed. 1992f. *The Irish World Wide: History, Heritage, Identity: Vol. 6, The Meaning of the Famine*. Leicester: Leicester University Press.
- Tagliamonte, Sali A. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
- 2012. *Variationist Sociolinguistics: Change, Observation, Interpretation*. Oxford: Wiley-Blackwell.
- Tieken-Boon van Ostade, Ingrid. 2009. *An Introduction to Late Modern English*. Edinburgh: Edinburgh University Press.
- Thomas, William I., and Florian Znaniecki. 1958. *The Polish Peasant in America, Original edn, 1918–1920*. New York: Dover Publications.
- Valdés, Mario J. 1996. La intrahistorio de unamuno y la nueva historio. *Revista Canadiense de Estudios Hispánicos* 21(1): 237–250.
- Vasconcelos, José Leite de. 1890. Dialectos alentejanos: contribuições para o estudo da dialectologia portuguesa. *Revista Lusitana* 2: 15–45.

Websites, Software and Online Resources

- AntConc*: Anthony, Laurence. 2015. *AntConc* (Version 3.4.4) Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net/software/antcon> (accessed 14 July 2014).
- Corpus Presenter*: Hickey, R. 2014. *Corpus Presenter* (Version 14). Essen: University of Duisberg-Essen. <https://www.uni-due.de/CP> (accessed 11 March 2015).
- DEM: Digitising Experiences of Migration*. <http://lettersofmigration.blogspot.fr> (accessed 30 June 2014).

- DIPPAM: Documenting Ireland: Parliament, People and Migration.* <http://www.dippam.ac.uk> and <http://gtr.rcuk.ac.uk/project/12B979D5-D08B-4B5B-92A5-58B5B292A587> (accessed 3 July 2014).
- ENROLLER: An Enhanced Repository for Language and Literature Researchers.* <https://enroller.nesc.gla.ac.uk> (accessed 30 June 2014).
- Fifteenth Literature of Irish Exile Autumn School* http://www.qub.ac.uk/cms/events/15th_LIE_2014/LIE_Oct_2014.htm (accessed 31 October 2014).
- MBDNL: Múin Béarla do na Leanbháin ('Teach the Children English')* <http://research.ncl.ac.uk/ni-language-migration> (accessed 3 July 2014).
- NAACE: The iPad as a Tool for Education* <http://www.naace.co.uk/publications/longfieldipadresearch> (accessed 24 April 2015).
- National Curriculum for England:* <https://www.gov.uk/government/collections/national-curriculum> (accessed 5 September 2014)
- O'Leary, Niall. 2014. Developing visualization tools for correspondence corpora. <http://development.nialloleary.ie/correspondence/correspondence.php> (accessed 5 September 2014).
- OxGarage:* <http://www.tei-c.org/oxgarage> (accessed 5 September 2014).
- Sankoff, David, Sali A. Tagliamonte and Eric Smith. 2012. *Goldvarb Lion: A multivariate analysis application for Macintosh.* Toronto/Ottawa: University of Toronto and University of Ottawa. <http://individual.utoronto.ca/tagliamonte/goldvarb.html> (accessed 9 August 2015).
- Stadler, Peter. 2013. Introduction to TEI. Presentation at Workshop 1 of the DEM Project. [29–30 May, Utrecht University]. <https://prezi.com/d2vahzuekxtx/introduction-to-tei> (accessed 31 October 2014).
- TEI: Text Encoding Initiative.* <http://www.tei-c.org> (accessed 19 May 2014).
- TEI SIG-Correspondence:* Text Encoding Initiative Special Interest Group-Correspondence. <http://wiki.tei-c.org/index.php/SIG:Correspondence> (accessed 5 September 2014).
- W3C: World Wide Web Consortium—XML Technology section* <http://www.w3.org/standards/xml> (accessed 30 October 2014).
- Wordsmith:* Scott, Mike. 2012. *WordSmith Tools version 6*, Stroud: Lexical Analysis Software. <http://www.lexically.net/wordsmith> (accessed 11 March 2015).

3

Engaging Users of Scottish Online Language Resources

Wendy Anderson and Carole Hough

1 Introduction

Publications on digital resources in the arts and humanities rarely focus on the users of such resources. Where they do, the users in question tend to be scholars. Indeed, the profile of even these users was for a long time poorly understood: Warwick (2012: 2) discusses the slow uptake in digital resources in the humanities at the end of the twentieth century, and reflects on the difficulties of establishing why users did not adopt digital resources in the manner and number that resource creators anticipated. Moreover, she notes that for a time, '[f]unding bodies also supported digital resources for humanities scholars, with little thought to, or predictions about, levels of possible use because they did not know how such predictions might be made' (Warwick 2012: 2). Resource use by the wider community is arguably even less well understood. However, with the increased attention paid in recent years by UK universities and funding bodies to public engagement and knowledge exchange, driven

W. Anderson (✉) • C. Hough
University of Glasgow, Glasgow, UK

by government priorities, projects are now obliged to interact more fully with user groups beyond academia and often to set out in considerable detail their plans for engagement and impact at the time of the funding application.¹ While we are still in the early days of measuring these phenomena reliably in the humanities disciplines, this means that funding bodies can at least be confident that researchers have fully considered the potential for the broader impact of their research. Nevertheless, the means of engaging with wider audiences often involve new technologies whose potential—and associated problems—are not yet fully understood. In addition, the most appropriate ways of engaging to the mutual benefit of project and user are likely to vary from project to project: there is as yet no tried-and-tested template to follow, and perhaps never will be.

This chapter presents an overview of several of the language and linguistics projects which have been developed in recent years in the English Language subject area at the University of Glasgow, and concentrates in particular on the ways in which these projects have sought to establish connections with the wider community and are continuing to engage with users, including researchers, school pupils, and members of the public. English Language at Glasgow has a strong tradition in empirical and textual research into language, in its contemporary forms as well as in its historical varieties, and in resource creation: new technologies are increasingly enabling us to exploit community connections at different stages of the creation and completion of linguistic resources.

First, we outline the four main projects discussed here: the *Scottish Corpus of Texts & Speech* (SCOTS); its sister project, the *Corpus of Modern Scottish Writing* (CMSW); *Scots Words and Place-names* (SWAP); and *Mapping Metaphor with the Historical Thesaurus*, touching also on the *Historical Thesaurus* itself, which forms the data for this last project. Because of their very distinct aims and requirements, the four projects have adopted different but overlapping strategies for engaging with users. Subsequently, we draw out a number of themes which have emerged in the creation and exploitation of some or all of these resources and

¹ See, for example, the AHRC's policy on impact, public engagement and knowledge exchange: <http://www.ahrc.ac.uk/What-We-Do/Strengthen-research-impact/Pages/Strengthen-Research-Impact.aspx> (last accessed 18 May 2014).

others like them. These range from identifying the most appropriate means of establishing and maintaining contact with wider audiences, to the crucial issues of the provision of metadata and sustainability of resources. Even in the decade or so in which most of the activity of these projects has taken place, the opportunities for engaging with users have changed—and increased—quite significantly. Our experience of all of these projects leads us to agree with Purnell et al. (2013: 403), who say that ‘[o]ne luxury that we as linguists have is that almost everyone likes to talk about how people talk, both their own speech and that of others’.

2 Language and Linguistic Projects at Glasgow

The four projects outlined in this section represent only a small set of the linguistic research projects that have been carried out in English Language at the University of Glasgow in recent years. We work in a unit that has particular strengths also in sociolinguistics, (socio-)phonetics, narrative, and Old, Medieval and Early-Modern English, with many projects involving knowledge exchange to a greater or lesser degree. These four have been selected to illustrate some of the major issues in relation to the nature and role of public engagement in resource creation and exploitation.

2.1 The *Scottish Corpus of Texts & Speech*

In addition to the tradition of empirical and textual research in English noted in Sect. 1, English Language at the University of Glasgow has a long-standing research interest in the continuum of language varieties stretching from Scottish English (or Scottish Standard English, the variety of English spoken in Scotland) to Scots (also known as Lowland Scots or Broad Scots, and often treated as a distinct Germanic language variety) (Corbett et al. 2003; Smith 2012; on the nature of the continuum of varieties, see Aitken and McArthur 1979). These two interests were brought together and consolidated around the turn of

the millennium in the form of the *Scottish Corpus of Texts & Speech* project. SCOTS created an online linguistic corpus of texts in Scots and Scottish English dating from 1945 to the present day.² In terms of its timing, the project fitted well in a period of significant change in Scotland, with the establishment of the Scottish Parliament in 1999, and the subsequent devolution of various powers. While it is not possible to measure precisely the effect that the heightened awareness of issues of Scottish identity had on people's engagement with the project, certainly we encountered a public that could see a genuine value in the linguistic resource we were creating.

The SCOTS project identified a gap in the available resources and a research requirement for a corpus of Scottish language varieties, in their various forms. Existing major British English corpora, such as the *British National Corpus* and the Bank of English, contained small quantities of Scottish material, but did not collect it systematically or in a balanced way. Information was also lacking on how extensively varieties of Scots were used and in what contexts, and on the relationship between current-day standard and non-standard languages in Scotland. Similarly, there was a requirement for up-to-date information on the linguistic features of modern Scots and Scottish English. As Wolfram et al. (2008: 1123) explain, '[i]t is essential to include different social, regional, and community voices, and to allow communities to speak for themselves'. SCOTS sought to do just that.³

SCOTS began in 2001, and was funded first by the Engineering and Physical Sciences Research Council and then by an Arts and Humanities Research Board (now Arts and Humanities Research Council) resource enhancement grant. The project formally ended in 2007, but texts have continued to be added to the corpus as they have become available, and

²The SCOTS project team was led by Professor John Corbett. Details of the full team and further information about the project are available on the website (for this and other resources mentioned here, see the websites and online resources list at the end of this chapter).

³Wolfram et al. (2008: 1115) make a further point which is highly relevant here: 'Communities that have been socialized into believing that their language variety is nothing more than "bad speech" are not particularly eager to celebrate this presumed linguistic inferiority, presenting a significant obstacle for the development of dialect awareness programs that celebrate local linguistic themes'. In the course of compiling the SCOTS corpus, we encountered divergent attitudes to—and indeed levels of awareness of—Scots as a language variety.

development work is still ongoing on the website. At the time of writing in May 2014, the resource contains over 1300 separate texts, comprising nearly 4.6 million words of text, of which 77 per cent is written language and 23 per cent spoken. The latter is accompanied by online audio (and, in a smaller number of cases, video) recordings and aligned orthographic transcriptions (on techniques of transcription alignment, see Anderson and Beavan 2005). The texts included exemplify a diverse range of genres, from prose fiction and records of the Scottish Parliament to semi-structured interviews and spontaneous speech. The online Advanced Search facility offers a concordancer and searches based on a large number of metadata categories, and has been integrated with Google Maps to allow users to identify texts based on the place of birth or residence of the author or speaker, where this information could be made available (see Sect. 3.3 below on metadata, and also Fig. 3.1). Users can identify where words or phrases are used, and investigate basic statistical information such as frequency for personal research and educational applications. Users can also freely download full texts or indeed the complete corpus, which enables analysis with proprietary corpus software. The resource therefore offers to researchers, educators and interested members of the public flexible access to a unique, sizeable and very varied collection of texts in Scots and Scottish English.

The success of SCOTS depended from the outset on the interest and involvement of the wider community. Unlike many corpus projects, which have an initial set of texts to digitize, or a schema of text types to be carefully represented in the corpus, SCOTS was obliged to establish the nature of the textual population from which it was sampling as the project progressed. That is, the focus of the project was as much on assessing the availability and nature of suitable texts as it was on digitization. For example, we knew that we would have quite easy access to large quantities of prose fiction in Scots, and felt sure that there would be a lot of personal correspondence in varieties of Scots too (whether or not we could obtain the permissions to use it), but we were less confident about finding examples of more peripheral genres. In the event, we found a wide range, including invoices, newsletters, and—very unexpectedly—a wedding ceremony in Scots, and many other genres in Scottish English too.

Your search profile

Criteria	Selection
Word search:Word/phrase (concordance)	dour <input type="checkbox"/>

Search results

65 occurrences of *dour*.
29 documents matching criteria (2.2% of corpus).
418,656 total words in matching documents (9.13% of corpus).

Show me: Map | Flags Residence | Concordance | Document list

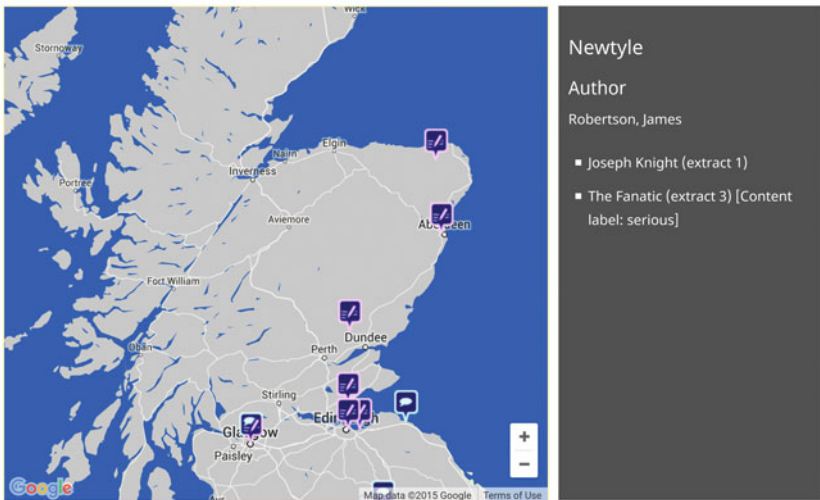


Fig. 3.1 Screenshot of SCOTS corpus, showing a search for the word-form *dour*, with associated Google Maps results

As a result of our somewhat opportunistic approach to data collection, the project team drew heavily on material supplied by the public in corpus compilation. This involved using various media to appeal for donations of texts. One of the most productive methods, in our experience, was using press releases issued by the university’s communications office at the time of the launch of the initial version of the online resource in 2004, which were then picked up by broadcast and print media. We also published pieces outlining the resource and seeking donations in Scottish

general-interest magazines with broad circulations, *ScotLit* (Anderson 2006) and *Leopard* (Anderson 2005). A sizeable proportion of the written texts in the corpus were obtained in response to these activities.

Much of the spoken text collection was also acquired through interactions with the public, often through more targeted approaches by the project team and associated students who would record members of their families, friends and Scottish public figures. In other cases, the engagement was less direct, with the SCOTS team being offered collections of recordings gathered by researchers for other projects and then returning to the participants themselves to seek the necessary permissions.⁴ This was the case, for example, for the substantial set of recordings of spontaneous interactions between caregivers and pre-school children in Buckie in the north-east of Scotland, a collection established by Jennifer Smith as part of a project on dialect acquisition (see, for example, Smith et al. 2009), and for the set of semi-structured discussions collected for the *BBC Voices* project (see the website listed at the end of this chapter, and also Upton and Davies 2013). In short, we were highly reliant on others, a risky strategy but one that paid off in terms of the range of usable textual material we were able to locate.

2.2 The Corpus of Modern Scottish Writing

The *Corpus of Modern Scottish Writing* (CMSW) was planned as a sister resource to SCOTS, and was funded by the AHRC from 2007 to 2010.⁵ Unlike SCOTS, which had been funded by a resource enhancement grant, CMSW was funded by a standard responsive-mode research grant. As Terras comments (2012: 59), '[w]hereas the 1990s were the "decade of digitization", we are now in the decade of digital belt-tightening, self-reflection and honest assessment of achievements in using digitized content within the humanities'. The design of the CMSW project therefore incorporated research outputs as well as resource creation: the

⁴ Other sets of recordings that we were offered we sadly had to decline because it was not possible to obtain permission for their inclusion in a freely available online corpus.

⁵ The CMSW project team was led by Professor John Corbett and the main researcher was Dr Jennifer Bann. Further details can be found on the project website (see section in References).

principal output is a book tracing the roots and development of literary Scots from the Older Scots period through to the twentieth century, focusing on orthographic strategies for spelling literary Scots (Bann and Corbett 2014).

While SCOTS had identified and filled a geographical gap in corpus provision, the principal aim of CMSW was to address the chronological gap that the creation of SCOTS had highlighted. The *Helsinki Corpus of Older Scots* represented the period from 1450 to 1700 (Meurman-Solin 1995) and SCOTS picked up the baton from 1700 to the present. But that left the period from 1700 to 1945, a very significant period both in terms of sheer length and also in terms of the changing influences on Scots and the development of the language, with 1700 being the date from which Modern Scots is conventionally dated. The period begins with the last stages of the standardization of written English, and takes in the ‘Vernacular Revival’ in literary Scots that produced writers like Robert Burns. CMSW was designed to provide evidence for the empirical analysis of these varieties in this period of time and to open up this evidence base for wider use. Its completion, in 2011, means that it is now possible to carry out empirical research into nearly six centuries of Scots. CMSW opens up numerous avenues for research into the forms and role of the Scots language, and interaction between Scots and Standard English, not to mention aspects of the history, culture and society of Scotland in the long period in question.

CMSW, which currently totals 5.5 million words, contains texts grouped into nine umbrella genres: administrative prose, expository prose, personal writing, instructional prose, religious prose, verse and drama, imaginative prose, and journalism. It also contains a collection of texts by eighteenth- and nineteenth-century ‘orthoepists’ or commentators on language, which, like all of the other genre groupings, but perhaps for more obvious reasons than the others, may be isolated and searched as a subcorpus in its own right. Researchers and other users can therefore compare the orthoepists’ pronouncements and recommendations for how the language should be spoken to the language in actual use at the time, and trace language change in light of this. Beal (2012) offers an overview of the respective roles of direct evidence (for example, orthoepists’ and grammarians’ statements on language) and indirect

evidence (including the evidence from authentic texts), specifically for our knowledge of earlier forms of pronunciation, but this can be broadened to other aspects of language such as grammar.

Unlike SCOTS, which drew heavily on contributions of texts from the public, the historical focus of CMSW required the team to rely instead on manuscript and print texts acquired from organizations such as the John Murray Archive of the National Library of Scotland, the Mitchell Library, Glasgow University Archives and Glasgow University Library Special Collections. All texts in the corpus are downloadable and are accompanied by high-quality images, which enables users to consult the original in cases where, for example, handwriting and layout are relevant for the research, or where there is uncertainty over the transcription.

Although the project team did not rely on textual contributions from the public to form the content of the corpus, engagement with a broader community was still one of the major aims of the project. This was achieved through a number of means. Early on in the project, the transcription and associated images of one particular text were released: this was the so-called ‘Kilmarnock Edition’ of Robert Burns’ *Poems, Chiefly in the Scottish Dialect* (1786), timed to mark the 250th anniversary of Burns’ birth, in 2009. A number of public talks were held as part of Glasgow’s West End Festival, an annual event in the local calendar, attracting capacity audiences. Finally, a blog was launched, drawing on the travel diary of the emigrant Thomas Crawford and charting—in ‘real time’ but 185 years later—his journey by sea from Scotland to Australia in 1825 (see Fig. 3.2).

Together, SCOTS and CMSW offer free access to over 10 million words of text in the continuum of language varieties from Scots to Scottish English. This allows everyone with an interest in these language varieties to investigate and teach them in new ways. We know that the resources are being used in university-level teaching, in the UK, Europe and beyond, by other researchers, and by lexicographers involved in the production of dictionaries of Scots and English. We also know that they are being used by members of the general public, as individuals still email us at the project address on a regular basis with queries and suggestions for texts and, occasionally, to express their appreciation for the website. It is difficult, however, to gauge the exact nature of the use of

Thomas Crawford's Diary, 1825

The screenshot displays a CMSW website interface. At the top, there is a navigation bar with the title 'Thomas Crawford's Journey'. Below this is a map showing a blue line representing the journey route across the United Kingdom and into the North Sea. The map includes labels for 'Edinburgh or Glasgow', 'United Kingdom', 'North Sea', 'Denmark', and 'Copenhagen'. Below the map, there is a search bar and a section titled 'The diary' with a brief introduction to the diary's content.

Uncategorized — Comments Off

09 Went to the Scotch Kirk in the forenoon
OCT 19

Sunday 9th. Octor. I went to the Scotch Kirk in the forenoon, and there was no Sermon in the afternoon till 1/2 past 6, when I intended to have gave to the Church, but I could not get on shore.

Uncategorized — Comments Off

PAGES CATEGORIES RSS

VIEW THOMAS CRAWFORD'S JOURNEY IN A LARGER MAP

Map data ©2015 GeoBasis-DE/BNK (©2009), Google, Inst. Geog. Nacional, Terms, 200 km

Google My Maps

Edinburgh or Glasgow

United Kingdom

North Sea

Denmark

Copenhagen

Hamburg

Isle of Man

SEARCH

The diary

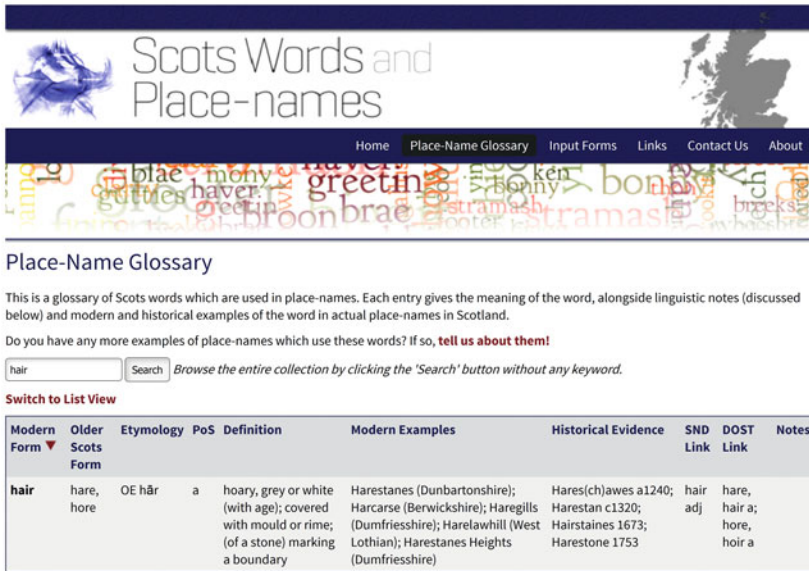
This is the diary of Thomas Crawford, a young man emigrating to Australia from Scotland in 1825. The long journey by sea took five months; we'll be putting his diary entries up on the dates he recorded, albeit 185 years later. The map at the top of this page shows his

Fig. 3.2 Screenshot from CMSW website of real time blog of Thomas Crawford's journey from Scotland to Australia

the resources by this latter group: we can simply assume that the number of people who use the resource is somewhat larger than the number who actually contact us.

2.3 Scots Words and Place-names

Another focus of research in Glasgow is Name Studies, an interest shared by English Language and the Celtic and Gaelic subject area. The place-names of Scotland derive from a range of historical languages including Brittonic, (Scots) Gaelic, Old English, Old Norse and Pictish, and comprise a major source of evidence for those languages and their speakers. Many later names, however, derive from Scots, and these have tended to receive less attention despite being equally rich in evidential value. One problem is that they are sometimes difficult to access, as they include informal names used within local communities but not represented on



The screenshot shows the website header with the title "Scots Words and Place-names" and a navigation menu. Below the menu is a word cloud. The main content area is titled "Place-Name Glossary" and contains a search bar with the word "hair" entered. Below the search bar is a "Switch to List View" button and a table with the following data:

Modern Form ▼	Older Scots Form	Etymology	PoS	Definition	Modern Examples	Historical Evidence	SND Link	DOST Link	Notes
hair	hare, hore	OE hār	a	hoary, grey or white (with age); covered with mould or rime; (of a stone) marking a boundary	Harestanes (Dunbartonshire); Harcarse (Berwickshire); Haregills (Dumfriesshire); Harelawhill (West Lothian); Harestanes Heights (Dumfriesshire)	Hares(ch)jaws a1240; Harestan c1320; Hairstaines 1673; Harestone 1753	hair	hare, hair a; hore, hoir a	

Glossary compiled by Dr Alison Grant of Scottish Language Dictionaries and the Scottish Place-Name Society.

Fig. 3.3 Screenshot of *Scots Words and Place-names* website, showing sample item *hair* in the place-name glossary

maps or in written documents. The *Scots Words and Place-names* project was designed to investigate the potential of social media for engaging the public in the exploration of such names, and of other regional uses of Scots terms (see Fig. 3.3).⁶

While in many areas of language study, as Wolfram et al. (2008: 1111) note, '[t]he specialized expertise of linguists sets up an asymmetrical relationship of authority with respect to language matters', place-name research represents a more equal partnership between academics and the wider public. As in sociolinguistics, it is important to develop good working relationships with participants, and indeed the branch of name studies known as socio-onomastics draws on many techniques from sociolinguistics (see Lieberman 1984; Leslie and Skipper 1990).

⁶ The SWAP website has now been archived (see section in references). Further information is available online in the Final Report (Hough et al. 2011), and in Bramwell and Hough (2014).

Palaeographical and philological training is required to trace place-name origins through the early spellings preserved in documentary and other sources, but local knowledge often proves crucial to interpretation. In the Preface to the first volume of the *Scottish Place-name Survey*, the authors describe how they ‘chapped on many doors, and spoke with many people regarding pronunciations, field-names and other aspects of local place-nomenclature’ (Taylor with Márkus 2006: ix). This rewarding but time-consuming process is replicated wherever serious survey work is undertaken. Through the use of social media, it may be possible to access local knowledge more efficiently, crowdsourcing in order to draw on a much larger pool of informants.

SWAP ran from March to November 2011 and was funded by JISC (formerly the Joint Information Systems Committee) under the Enriching and Developing Community Content programme. It was a collaboration between the University of Glasgow, Scottish Language Dictionaries (SLD) and the Scottish Place-Name Society (SPNS). While the SPNS is the main organization for research into place-names in Scotland, SLD is the national body for lexicography, maintaining and updating the online *Dictionary of the Scots Language (DSL)* alongside print dictionaries directed towards different user groups. Material for updating the dictionaries with new words, and with additional information on the meanings, spellings, and chronological and geographical range of existing words, is held in the Word Collection database, which draws on a range of sources including corpora and contributions from local informants (see Robinson 2013). Again, social media may prove a valuable tool for this kind of enterprise. Like other major dictionaries including the *OED*, SLD is making increasing use of place-name data, a type of source largely overlooked by early lexicographers (see Scott 2004). There was therefore a strong rationale for combining research which crowdsourced Scots place-names with research which crowdsourced Scots words.

SWAP was spread across four online platforms: a public website, Facebook, Twitter and Glow, the Scottish schools’ intranet. The website hosted a message board together with information about the project, input forms for submitting Scots words and place-names, and a glossary of place-name elements which was built up gradually during the lifetime

of the project.⁷ The Facebook and Twitter accounts were linked to the website, and were used to initiate discussions, appeal for information and disseminate news items. Contributions were solicited thematically, by posting a weekly topic on each of the three platforms. Examples included words for alcoholic drinks, place-names containing specific terms such as Scots *heid* 'head', and nicknames for people living in particular towns. For each topic, a selection of known examples from different parts of Scotland was provided, and the public were encouraged to contribute others from their own localities. These further examples were then fed into SLD's Word Collection, or into the glossary of place-name elements. The topics were chosen in consultation with the project partners so as to be maximally useful for ongoing research. In addition, place-name sources were trawled for the elements glossary, and a subcorpus of the Word Collection was fed through Wordnik and Google searches in order to find evidence from other blogs, message boards and websites for the usage of words and phrases that at that time lacked supporting contexts.

Glow served as the platform for a schools competition which ran as part of the project in order to engage younger age groups and raise the profile of Scots in primary and secondary education. Although not envisaged at the planning stages of SWAP, this was devised as a solution to the ethical problems presented by the fact that children are not allowed to use Facebook. Entries could relate to any aspect of the Scots language or place-names, and could be in any format: submissions included essays, poems, songs, short stories and illustrated booklets. Finalists from each of the five age groups were chosen by the judges, who included the novelists Amal Chatterjee and Louise Welsh. The winners were then decided by peer vote on Glow, and announced at the prize-giving in Glasgow hosted by the University Rector, Rt Hon. Charles Kennedy MP. As part of the event, the finalists, together with their parents and teachers, were given a tour of the university, including an exhibition on Scots created for the occasion by Robert MacLean of the University Library's Special Collections Department, entitled 'From 'Makaris' to Makars: Scots Literature in Special Collections' (see section in References).

⁷An element is an individual component of a place-name. For instance, the Scots element *law* 'round hill' occurs in a number of place-names including Lawhead, Lawmuir and Meikle Law.

The main outputs of the project were the exhibition, new material for SLD's Word Collection, and the glossary of Scots place-name elements created by Alison Grant of SLD and SPNS. The glossary was initially based on data from the electronic files of *DSL*, and was supplemented with additional material throughout the project. It now comprises the most comprehensive and authoritative source of information on Scots terms in place-names, and is available as a searchable database on the project website. Less tangible but equally valuable outcomes of SWAP were the more strategic use of social media by ourselves and our project partners and, more significantly, the raised profile of Scots in the community and in schools.

2.4 *Mapping Metaphor with the Historical Thesaurus*

The experiences of public engagement in all of the projects discussed so far are now shaping the research and engagement plans of a new project which is currently underway at Glasgow, *Mapping Metaphor with the Historical Thesaurus*.⁸ The project, funded by the AHRC from 2012 to 2015, exploits the data of the *Historical Thesaurus of English* (published as the *Historical Thesaurus of the Oxford English Dictionary*, Kay et al. 2009) to locate and analyse metaphorical transfer in the recorded vocabulary of English, from the Anglo-Saxon period to the present day. This is achieved through the automatic identification of lexical overlap between semantic categories and subsequent detailed manual analysis to identify cases where the lexical overlap reflects metaphorical transfer. As an illustration, the semantic categories of Excitement and Taste share lexical items because of a metaphorical connection by which speakers of English talk about exciting things in terms borrowed from the domain of taste (such as *savour*, *zesty*, *spicy*, *piquant*).

We plan to launch our main public output, the 'Metaphor Map', early in 2015 (see Fig. 3.4). This will be an interactive visual representation of all of the systematic metaphorical connections in English, browsable and searchable by semantic domain, keyword and, ultimately by time period.

⁸Information on the *Mapping Metaphor with the Historical Thesaurus* project can be found on the website (see section in References). Once complete, the 'Metaphor Map' resource will also be accessible from this site.

the eventual resource will look like. We expect that this use of social media will become more dialogic once the Metaphor Map is fully online. We have held public talks as part of Glasgow's West End Festival and Aberdeen's Elphinstone series, finding enthusiastic audiences including creative writers among a wider public. In addition, we have held events under the banner of 'Science Sunday' at the Glasgow Science Festival (where SWAP was also represented) and as part of 'Explorathon', run by the European Researchers' Night. The *Historical Thesaurus*, which was also created at Glasgow, itself achieved major impact on publication in 2009 and continues to do so, partly through its incorporation into the online *Oxford English Dictionary* (see Kay 2012): with *Mapping Metaphor*, we are beginning to build on this impact, tapping in to a similar general audience with a proven interest in language and words. We are also seeking to exploit the educational angle of the Metaphor Map resource through collaboration with secondary-school learners and teachers, and education officers; at the present time, however, this is still somewhat speculative.

3 Key Issues

We draw out here some of what we feel are the main issues which we encountered in carrying out these diverse projects and others like them. We have ordered these with a nod to typical chronology in the lifetime of a project, focusing on the four projects as appropriate to each set of issues. There is, however, a strong caveat that many of the earlier questions continue to be relevant throughout a project's duration, and indeed in some cases well beyond a funded stage, and that some of the later questions need to be considered and planned for from the outset.

3.1 Making Contact

In our experience, potential academic users of digital or other resources are generally easy to identify, and can be contacted through email lists and the like. Potential non-academic users are more diverse, and contacting them is correspondingly less straightforward. 'First catch your hare',

as early cookery books are reputed to have said.⁹ Any public engagement project has to strike a balance between covering a wide expanse of the terrain in the attempt to flush out hares, and focusing on selected areas where they are known to congregate. This tends to be achieved through a combination of media publicity and targeted approaches to relevant groups and societies. On the one hand, television, radio and press coverage reaches a very large cross-section of the population; on the other, there is no guarantee of a response, whereas members of an existing group are more likely to become actively involved, especially if contacted directly. Although the SWAP project, for example, was promoted through media publicity on programmes such as Radio Scotland's *Culture Café*, the involvement of our project partners was crucial in gaining access to interested user communities. SPNS, with around 375 members, provided a bedrock of support in different parts of the country, while SLD incorporated the project into its outreach programme, with very positive results. Further participants were recruited through the Scots Language Centre, others by the traditional means of distributing publicity materials at various events. An unexpected bonus of the schools competition was the publicity it generated for the wider project. Many of the participating schools featured SWAP on their home pages, and once the finalists were announced, coverage also appeared in local media outlets such as the *Eskdale and Liddesdale Advertiser*, *Hawick News*, *New Shetlander* and *The Shetland Times*.

We have often tried to tie projects in to wider events, to encourage interest and press coverage. Hence the SCOTS corpus was launched online on St Andrew's Day 2004, and, as noted above, CMSW released the text and images of Burns' Kilmarnock edition to coincide with the 250th anniversary of the Scottish national poet's birth. Even when there is no opportunity to coordinate with external events and anniversaries, in our view timing is still instrumental in successfully involving the public in academic projects. With SCOTS, the timing of different stages of the corpus compilation process also proved to be crucial. In order to attract interest in the project and to allow people to appreciate the nature of our research, it was important to be able to give a clear indication of the

⁹The phrase is apocryphal, the actual wording being 'First case your hare'.

finished resource. We therefore staggered our calls for donations of texts, holding back on general calls in the media until the time of the launch of the website with initial corpus texts. This meant that potential contributors could look online to see exactly what we planned to do with their texts, read about the background to the project, and better understand our research aims.

Similarly with SWAP, the glossary of place-name elements was revealed gradually on the website throughout the course of the project, so that people responding to calls for additional examples could see how they would contribute to the overall resource. *Mapping Metaphor* had been running for nearly 18 months before the first public engagement events took place, because here the focus was on presenting some of the initial results of the project in order to stimulate interest: this contrasts markedly with the strategy for engaging with academic users, which has been ongoing from the very outset of the project with a focus on methodology.¹⁰ Finding the most appropriate time to engage with the wider audience, whether to collect data or publicize project research, is closely connected to the findings of Cameron et al. (1997: 145), that ‘power relations are strongly affected by the methods we are constrained to adopt in “doing research”’.

Timing was also crucial to the success of the SWAP schools competition. Following consultation with teachers and with colleagues in the School of Education at the University of Glasgow, this was scheduled for the final weeks of the school year, when there is more flexibility in the curriculum. We were very pleased that, despite the short notice of the prize-giving following the judging process, finalists from all except one school came to Glasgow for the event, which was held in the University Library with support from the Friends of Glasgow University Library. In other respects, the short time frame of SWAP probably had a detrimental effect. Nine months proved less than fully adequate to establish a strong social media presence and grow a user community, and although the project had many enthusiastic participants, we felt that far more could have been achieved over a longer period.

¹⁰ Or rather, methodologies: for instance, presentations to digital humanities groups concentrate on the technical side of the project, while theoretical issues are foregrounded in presentations to linguists.

The later projects discussed here, SWAP and *Mapping Metaphor*, have both used Twitter to advantage in disseminating project information widely. In the early stages of SWAP and *Mapping Metaphor*, we followed and tweeted relevant Twitter sites, and asked colleagues and media figures with large Twitter followings to tweet about the projects. SWAP benefited particularly from tweets by Neil Oliver, the Scottish archaeologist and broadcaster, for example, whose followers included many interested members of the public. In general, we found Twitter invaluable for raising the profile of the projects and advertising events. For SWAP, Facebook was more useful for the research itself, allowing for sustained discussions and fuller data (see Hough et al. 2011). One of our project partners, SPNS, set up a Twitter account as a result of taking part in SWAP; more recently, it has also set up a Facebook page. *Mapping Metaphor* does not currently have a presence on Facebook, but instead uses a WordPress blog for similar purposes.¹¹

3.2 Maintaining Contact

Having caught your hare, it is necessary to preserve it until ready for the pot. Even after a project has attracted their attention, users, whether academic or not, will quickly lose interest unless the momentum is maintained. Keeping the public engaged is no less important than making contact in the first place, and requires a different range of strategies.

Here there is a crucial difference between projects to which the public are asked to contribute and those which they are simply invited to use. At one end of the spectrum, SCOTS and SWAP depended fundamentally on contributions from the public; at the other, there is little if any opportunity for public input into the creation of the *Mapping Metaphor* resource (although we will include a mechanism for users to convey their views on the resource and comment on any errors or alternative interpretations). CMSW falls somewhere between the two poles. It is in general easier to acquire feedback when users have been involved from the outset, so both SCOTS and SWAP were able to respond flexibly to the changing

¹¹ See Ross (2012) for an overview of social media including microblogging in academia, and as a means of engaging with the wider community.

dynamics of the situation. For SCOTS, we quickly learned that it would normally be difficult to reuse existing recordings of speech, so focused largely on soliciting donations of written texts from the public, while establishing procedures for making new recordings ourselves. As regards SWAP, we found that not only did appeals for informal place-names result in a greater quantity of responses than appeals for official names, but that the quality of the data was better, so we began to focus increasingly on this area of the research. Similarly, although the public responded enthusiastically to appeals for local terms relating to individual topics, many of the words submitted were already well documented and therefore of little research interest; on the other hand, focused appeals for information on specific words for which we had identified gaps in the record resulted in some valuable data. Again, therefore, we were able to adapt our approach during the course of the project. This kind of iterative dialogue is not so practicable for a project like *Mapping Metaphor*, where the role of the non-academic community is largely limited to that of users of the end product, so it is all the more important to gain feedback through other means. Among them is the involvement of a variety of focus groups in testing early versions of the Metaphor Map in order to ensure that it will be optimally useful to a wide range of likely users.

Aside from the issue of feedback, public involvement in the creation of a resource may have implications for user engagement. In theory, people are more likely to use a resource if they have helped to create it. Unfortunately it is difficult to test out that theory. SWAP was envisaged as a community collection building exercise, with the public developing a sense of ownership that would lead to continuing engagement with the resource after the close of the funding period. To what extent that expectation was fulfilled is uncertain. Some of our Facebook and Twitter followers were also contributors to the website, but the statistics are difficult to ascertain, and we did not have a mechanism in place for tracking individual users through different stages of the project. Moreover, as the submissions still being received via the website are anonymous, with the online input forms simply recording optional information on age, background, gender and occupation, we have no way of knowing whether the contributors were involved during the project itself or became aware of it more recently. Similarly, the proportion of contributors to SCOTS who are also users of

the corpus is unknown. Empirical research in this area is badly needed, and would help significantly in the planning of future projects.

Again it is important to bear in mind that there are different types of non-academic users, with varying needs and interests. Focus groups with teachers, held by the *Mapping Metaphor* project in summer 2014, were aimed partly at gaining feedback on the prototype Metaphor Map resource, and partly at developing associated teaching tools. The teaching potential of the *Historical Thesaurus* was established through online packages created by two previous projects, both funded by the Higher Education Academy English Subject Centre (Hough and Kay 2007; Kay and Corbett 2008). Aimed at undergraduate students as well as lifelong learners, they have received appreciative feedback from as far afield as Florida. Both the primary and secondary education sectors have also proved receptive to academic research, and indeed we found that primary schools engaged particularly enthusiastically with SWAP, possibly because there are fewer constraints on the curriculum for pupils of that level. In order to promote the SWAP schools competition, we produced a selection of teaching materials which were made available on the Glow site. A number of teachers commented favourably on these, and they had clearly also been used by others, as in some instances entire classes submitted entries on related topics. The appetite shown for such materials has led to further initiatives, including a place-name resource for schools developed by members of the AHRC-funded Scottish Toponymy in Transition project team at the University of Glasgow in collaboration with Education Scotland. This is now available within the *Studying Scotland* website.

3.3 Supporting Users in Resource Exploitation

Key to maintaining users' engagement in online resources is the provision of appropriate support. This can take the form of straightforward information in accessible, non-specialist language, available alongside the online resource itself, as with SCOTS and CMSW, and as is planned for *Mapping Metaphor*. This applies of course to public engagement, and likewise to students and other scholars, whose use of resources can also be

encouraged and supported by textbooks and scholarly writing: SCOTS, for example, features quite heavily in a practical book for students on research techniques for exploring online corpora written by two of the members of that project team (Anderson and Corbett 2009).

Metadata is another important aspect in the provision of a resource that users can exploit as fully as possible. As Burnard (2004, n.p.) states, 'it is no exaggeration to say that without metadata, corpus linguistics would be virtually impossible'. He goes on to explain this by means of an analogy:

A typical corpus analysis will therefore gather together many examples of linguistic usage, each taken out of the context in which it originally occurred, like a laboratory specimen. Metadata can restore that context by supplying information about it, thus enabling us to relate the specimen to its original habitat. [...] Without metadata, the investigator has nothing but disconnected words of unknowable provenance or authenticity. (Burnard 2004, n.p.)

While metadata is perhaps more crucial for academic users, nevertheless, appropriate metadata relating to the language contained in a resource is also valuable for general users. The two corpus projects discussed here, SCOTS and CMSW, both incorporate detailed textual and personal metadata, and we know from email correspondence from general users that they also tap into this para-textual information to help them to contextualize the data. Details of dates and the birthplaces of speakers appear, perhaps not surprisingly, to be the most sought-after pieces of information.

This data was not straightforward to gather, and required a considerable input of time and effort on the part of the project research assistants in particular. A large set of permissions forms and associated permissions tracking in the administrative database was required for SCOTS, to acquire the relevant rights to include the (written, audio and video) texts in the corpus and the desired personal and textual metadata alongside them. The permissions forms were unavoidable, but will certainly have discouraged some potential contributors from donating their texts to the project. Texts could not of course go into the online corpus without the relevant permissions being granted; however, personal metadata was

optional. This means that the fullness of the personal metadata is variable across the corpus. One of the challenges of the SWAP project, similarly, was that it was difficult to combine lively, informal discussion on the website and on Facebook with legalistic requests for personal information. A compromise was for such information to be collected through the input forms, resulting in different levels of metadata for material that had been harvested in different ways. For CMSW, the problems of acquiring permissions and metadata were lessened, as all but the latest printed texts in the corpus were already out of copyright, and basic information on most of the authors could be readily obtained from sources like the online *Oxford Dictionary of National Biography*.

Naturally, it is easier to deliver a valuable resource when the resource creators know something of the profile of the users. The mechanisms for collecting information about the users of online resources are more sophisticated now than they were in 2004 when SCOTS first launched. For that project, we were able through usage log data and emails to the dedicated contact address to establish informally a general profile of the types of user who were accessing the resource: this included users interested in Scottish culture, history and politics and a surprisingly large number of people interested in tracing their family history, alongside the more predictable groups of students, teachers and researchers in language and linguistics. While the majority of users were local, from Scotland and the UK, there was also a steady stream of users from further afield, including the United States, Canada, Europe, South America and Asia. We intend, with *Mapping Metaphor*, to use the much more sophisticated user analysis information that Google Analytics can provide: indeed we are already using this to enhance our understanding of the readers of our WordPress blog.

3.4 Resource Sustainability

A crucial issue for all creators of digital language resources, whether intended for general audiences or solely for academic audiences, is that of the mid- to long-term sustainability of the resource. This is a question that needs to be considered from the planning stages, as the demands on time and money are considerable and not fully predictable, given the

ever-changing technological landscape which we are constantly negotiating. The difficulties are aggravated by the current funding mechanisms, at least in the UK. As Terras (2012: 58) explains, '[f]ewer and fewer funding calls are emanating from funding councils, and very few calls exist to provide continuation funding for established projects'. Unless resources have a deliberate finite lifespan, researchers therefore need to be creative in the ways in which they manage the continued existence, and ideally ongoing development, of their resources. In our experience with the projects discussed here, there are various ways in which at least a mid-term future for individual resources can be found.

The ideal solution is perhaps for the full cost of the ongoing maintenance of resources to be absorbed by the institution hosting them: in practice, this is increasingly uncommon. Where there is continuity of staff, especially on the technical side, proportions of costs can be absorbed by allocating portions of staff time to resource maintenance, alongside the creation of new funded resources. Where new projects have explicit links to existing ones, the creation of a new resource may entail maintenance or development work on the existing one. To a certain extent, this was the case for CMSW, following on from SCOTS and hosted on the same website, and for *Mapping Metaphor*, which exploits the *Historical Thesaurus of English* and has been a factor in ongoing development of that parent resource. SWAP was by far the shortest of the four projects, so sustainability was a key issue from the outset. The place-name glossary stands as a permanent legacy of the research, as does the online exhibition. The input forms for submitting Scots words and place-names also remain in use, and continue to attract contributions. While the generated data continues to inform the Scottish Language Dictionaries' collections, there has been no backwash effect to further develop the SWAP resource.

A further way in which researchers can encourage the longevity of their resource is to exploit opportunities for their incorporation into meta-resources. An early proof-of-concept project aiming to create an integrated online repository to enhance linguistic resources was ENROLLER, which was funded by JISC and ran from 2009 to 2011. A frank account of the challenges involved can be found in Anderson (2013).

Although difficult to quantify, sustainability also encompasses the wider continuing impact of the projects. SCOTS and CMSW are now

exploited as sources of lexical information by lexicographers of Scots and English, and their texts are selected for use in Higher and A-level examinations by national examination boards. For SWAP, alongside our weekly theme, some of our Facebook followers began their own off-topic discussions, so that the project began to develop a life of its own. Also, some of the schools that participated in the SWAP competition went on to undertake their own projects. SWAP has maintained momentum since the formal close of the project through participation in various external initiatives. These include a partnership activity on place-names within the BBC's *The Great British Story: A People's History* roadshow at the Riverside Museum in Glasgow in June 2012, and a workshop on Scots and place-names at YouthLink Scotland's *Digitally Agile Community Learning and Development National Stakeholders Event* at Dynamic Earth in Edinburgh in January 2013.

Other mechanisms involve continued close collaboration with project partners. Of the four projects outlined here, this was most pivotal to SWAP, not only in providing research material and advice but in helping to establish links with interested user groups such as the c.375 members of SPNS, many of whom are non-academics. SLD's Education and Outreach Officer, Elaine Webster, played a particularly important role in liaising with schools, building on her existing contacts to promote the SWAP competition. She has continued to be involved in follow-up initiatives such as the online place-names resource for schools mentioned in Sect. 3.2. Such productive connections with individuals are, in our experience so far, tremendously important to the success of projects with a longer-term public engagement or knowledge exchange component.

4 Conclusion

We know from experience that there is considerable scope for online linguistic resources to have an impact on the general public when these resources are well publicized, free of charge, and supported by appropriate usage guidance such as online help manuals, and when they incorporate a means of dialogue with the research team, whether that be a blog or Twitter feed, or simply an email address or web form. It goes without

saying, perhaps, that different projects will benefit from different forms of public engagement. Certainly that has been our experience in the context of the projects discussed here. While SCOTS and SWAP relied heavily upon the wider community in the very creation of the respective resources, CMSW and *Mapping Metaphor* have looked elsewhere for data, and interacted with users principally at more advanced stages. We have also been able to take advantage of new mechanisms for engagement over time, especially in light of new social media such as Twitter and Facebook.

Rarely has the public engagement of the various projects described here taken exactly the form that was originally intended, and we believe this is normal. Some of the hares we started did not run; others turned out to be veritable Easter hares, bringing unanticipated benefits to the projects. Indeed this is as it should be. Like Wolfram et al. (2008: 1130), we would contend that '[i]n an important sense, engagement is more of a process than a final product' and therefore that 'we should not be surprised that our final product often bears only a faint resemblance to our original idea'. Rather, we have enjoyed taking opportunities as they have arisen, and we believe that the projects are all the more valuable for this.

References

Books and Articles

- Aitken, A.J., and Tom McArthur, eds. 1979. *Languages of Scotland*. Edinburgh: Chambers.
- Anderson, Jean. 2013. *Enroller*: an experiment in aggregating resources. In *Language in Scotland: Corpus-based Studies*, ed. Wendy Anderson, 273–294. Amsterdam and New York: Rodopi.
- Anderson, Wendy. 2005. Is there Doric in your attic? *Leopard* 319.
- . 2006. Your country needs YOUSE! *ScotLit* 34: 13–14.
- Anderson, Wendy and David Beavan. 2005. Internet delivery of time-synchronised multimedia: the SCOTS corpus. *Proceedings from the Corpus Linguistics Conference Series* 1(1). Available online: <http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx/>

- Anderson, Wendy, and John Corbett. 2009. *Exploring English with Online Corpora*. Basingstoke: Palgrave Macmillan.
- Bann, Jennifer, and John Corbett. 2014. *Scots Spelling: The Orthography of Literary Scots*. Edinburgh: Edinburgh University Press.
- Beal, Joan. 2012. Evidence from sources after 1500. In *The Oxford Handbook of the History of English*, eds. Terttu Nevalainen, and Elizabeth Closs Traugott, 63–77. Oxford: Oxford University Press.
- Bramwell, Ellen, and Carole Hough. 2014. Scots in the community: place-names and social networking. In *Names in Daily Life: Proceedings of the XXIV ICOS International Congress of Onomastic Sciences*, ed. Joan Tort, 294–303. Barcelona: Available online: www.gencat.cat/llengua/BTPL/ICOS2011 Generalitat de Catalunya.
- Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, Ben Rampton, and Kay Richardson. 1997. Ethics, advocacy and empowerment in researching language. In *Sociolinguistics*, eds. Nikolas Coupland, and Adam Jaworski, 145–162. Houndmills: Macmillan (Originally published in *Language and Communication* 13(2): 81–94 in 1993.)
- Corbett, John, J. Derrick McClure, and Jane Stuart-Smith, eds. 2003. *The Edinburgh Companion to Scots*. Edinburgh: Edinburgh University Press.
- Kay, Christian J. 2012. Developing *The Historical Thesaurus of the OED*. In *Current Methods in Historical Semantics*, eds. Kathryn Allan, and Justyna A. Robinson, 41–58. Berlin and Boston: De Gruyter Mouton.
- Kay, Christian, Jane Roberts, Michael Samuels, and Irené Wotherspoon, eds. 2009. *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press.
- Leslie, Paul L., and James K. Skipper. 1990. Towards a theory of nicknames: a case for socio-onomastics. *Names* 38: 273–282.
- Lieberson, Stanley. 1984. What's in a name? ... some sociolinguistic possibilities. *International Journal of the Sociology of Language* 45: 77–87.
- Meurman-Solin, Anneli. 1995. A new tool: The Helsinki Corpus of Older Scots (1450–1700). *ICAME Journal* 19: 49–62.
- Purnell, Tom, Eric Raimy, and Joseph Salmons. 2013. Making linguistics matter: building on the public's interest in language. *Language and Linguistics Compass* 7(7): 398–407.
- Robinson, Christine. 2013. The use of corpora in lexicographical research in Scots. In *Language in Scotland: Corpus-based Studies*, ed. Wendy Anderson, 237–252. Amsterdam and New York: Rodopi.
- Ross, Claire. 2012. Social media for digital humanities and community engagement. In *Digital Humanities in Practice*, eds. Claire Warwick, Melissa Terras,

- and Julianne Nyhan, 23–45. London: Facet Publishing (in association with UCL Centre for Digital Humanities).
- Scott, Margaret. 2004. Uses of Scottish place-names as evidence in historical dictionaries. In *New Perspectives on English Historical Linguistics: Selected papers from 12 ICEHL, Glasgow, 21–26 August 2002*, Volume II: *Lexis and Transmission*, Christian Kay, Carole Hough and Irené Wotherspoon (eds), 213–224. Amsterdam and Philadelphia: John Benjamins.
- Smith, Jennifer, Mercedes Durham, and Liane Fortune. 2009. Universal and dialect-specific pathways of acquisition: caregivers, children, and t/d deletion. *Language Variation and Change* 21: 69–95.
- Smith, Jeremy J. 2012. *Older Scots: A Linguistic Reader*. Woodbridge, Suffolk: Boydell Press.
- Taylor, Simon with Gilbert Márkus. 2006. *The Place-names of Fife*. Volume 1: *West Fife between Leven and Forth*. Donington: Shaun Tyas.
- Terras, Melissa. 2012. Digitization and digital resources in the humanities. In *Digital Humanities in Practice*, eds. Claire Warwick, Melissa Terras, and Julianne Nyhan, 47–70. London: Facet Publishing (in association with UCL Centre for Digital Humanities).
- Upton, Clive, and Bethan L. Davies, eds. 2013. *Analysing 21st-century British English: Conceptual and Methodological Aspects of the BBC 'Voices' Project*. London: Routledge.
- Warwick, Claire. 2012. Studying users in digital humanities. In *Digital Humanities in Practice*, eds. Claire Warwick, Melissa Terras, and Julianne Nyhan, 1–21. London: Facet Publishing (in association with UCL Centre for Digital Humanities).
- Wolfram, Walt, Jeffrey Reaser, and Charlotte Vaughn. 2008. Operationalizing linguistic gratuity: from principle to practice. *Language and Linguistics Compass* 2(6): 1109–1134.

Websites and Online Resources

- Bank of English, and Collins Corpora*. <http://www.mycobuild.com/about-collins-corpus.aspx> (accessed 18 May 2014).
- BBC Voices*. <http://www.bbc.co.uk/voices> (accessed 18 May 2014).
- British National Corpus*. <http://www.natcorp.ox.ac.uk> (accessed 18 May 2014).
- Burnard, Lou. 2004. Metadata for corpus work. <http://users.ox.ac.uk/~lou/wip/metadata.html> (accessed 18 May 2014).

- Corpus of Modern Scottish Writing*. <http://www.scottishcorpus.ac.uk/cmsw> (accessed 18 May 2014).
- Dictionary of the Scots Language*. <http://www.dsl.ac.uk> (accessed 18 May 2014).
- ENROLLER: An Enhanced Repository for Language and Literature Researchers*. <http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/enroller> (accessed 18 May 2014).
- Friends of Glasgow University Library*. <http://www.gla.ac.uk/services/library/collections/friends> (accessed 18 May 2014).
- From 'Makarīs' to Makars: Scots Literature in Special Collections*. <http://www.gla.ac.uk/services/specialcollections/virtualexhibitions/frommakaristomakarsscotsliteratureinspecialcollections> (accessed 18 May 2014).
- Glasgow Science Festival*. <http://www.glasgowsciencefestival.org.uk> (accessed 18 May 2014).
- Google Analytics*. <http://www.google.com/analytics> (accessed 18 May 2014).
- Hough, Carole, Ellen Bramwell and Dorian Grieve. 2011. *JISC Final Report: Scots Words and Place-names*. JISC. <http://www.jisc.ac.uk/media/documents/programmes/digitisation/swapfinalreport.pdf> (accessed 18 May 2014).
- Hough, Carole and Christian Kay (eds). 2007. *Learning with the Online Thesaurus of Old English* (TOE) <http://oldenglishteaching.arts.gla.ac.uk/oeteach.html> (accessed 18 May 2014).
- Kay, Christian and John Corbett (eds). 2008. *Word Webs: Exploring English Vocabulary*. <http://www.gla.ac.uk/schools/critical/research/fundedresearch-projects/wordwebs> (accessed 18 May 2014).
- Mapping Metaphor with the Historical Thesaurus*. <http://www.glasgow.ac.uk/metaphor> (accessed 18 May 2014).
- Oxford Dictionary of National Biography*. <http://www.oxforddnb.com> (accessed 18 May 2014).
- Scottish Corpus of Texts & Speech*. <http://www.scottishcorpus.ac.uk> (accessed 18 May 2014).
- Scots Language Centre*. <http://www.scotslanguage.com> (accessed 18 May 2014).
- Scottish Place-name Society*. <http://www.spns.org.uk> (accessed 18 May 2014).
- Scots Words and Place-names*. <http://swap.nesc.gla.ac.uk> (accessed 18 May 2014).
- Scottish Language Dictionaries*. <http://www.scotsdictionaries.org.uk> (accessed 18 May 2014).
- Scottish Toponymy in Transition*. <http://www.glasgow.ac.uk/stit> (accessed 18 May 2014).
- Studying Scotland*. <http://www.educationscotland.gov.uk/studyingscotland/resourcesforlearning/learning/Contextsforstudy/placenames> (accessed 18 May 2014).

4

From Legacy Regional Language Materials to Public Engagement: The Interactive Online Dialect Atlas of Newfoundland and Labrador

Sandra Clarke

1 Introduction

Throughout the English-speaking world, there is huge public interest in regional speech differences, particularly differences of vocabulary. The ‘voracious public appetite for dialect maps’ (Zimmer 2013) has been

This project could not have been completed without generous funding from the Public Outreach Dissemination Grant Program of the Social Sciences and Humanities Council of Canada, along with Memorial University’s Institute of Social and Economic Research, its J.R. Smallwood Foundation, and its student research assistant programs. The Dialect Atlas also would not have been possible without the generous input of many; a full listing is provided at <http://www.dialectatlas.mun.ca/about/#aboutAcknowledgements>. I would especially like to signal out the Atlas design team from Memorial’s Distance Education, Learning and Teaching Support (DELTS) unit, as well as CCWebworks; my co-researchers Philip Hiscock and Robert Hollett; and Memorial’s English Language Research Centre manager Suzanne Power. I am very grateful to Gerry Porter of DELTS and David Cantwell of CCWebworks for providing me with technical information for this chapter. I am also extremely grateful to David Mercer of Memorial University Library’s Map Room for designing the two maps representing the communities that appear in our online Dialect Atlas (Fig. 4.1).

S. Clarke (✉)

Memorial University of Newfoundland, St. John’s, NL, Canada

fed by the enormous strides in computer technologies over the past two decades. These have yielded interactive websites inviting visitors to upload their responses to online lexical surveys and view the dynamic maps that result. Online maps showing self-reported regional differences in selected aspects of contemporary vocabulary are now available for many varieties of English, including those spoken in the United Kingdom (the Word Maps portion of the *BBC Voices* website), Australia (the *Australian Word Map*), North America (among them the *Harvard Dialect Survey*, Vaux and Golder 2003) and the wider English-speaking world (Vaux and Jøhndal's 2007 *Cambridge Online Survey of World Englishes*).¹ These maps range considerably in their degree of technical sophistication, from fairly simple tile displays (as on the *BBC Voices* site) to the heat-map visualizations in Katz's (2013) update of the *Harvard Dialect Survey*.²

Linguists with a primarily academic audience in mind have also availed themselves of recent technological developments, using a variety of different approaches, and targeting not only the traditional older rural speakers favoured by dialect geography, but also a full range of contemporary speakers of English. Viereck and Ramisch (1991, 1997) used computerized analytical techniques to produce a reworked print version of selected grammatical and lexical components of the Survey of English Dialects (compare the more traditional print atlases of Orton et al. 1978; Upton et al. 1987; Upton and Widdowson 1996). Kretzschmar's *Linguistic Atlas Projects* website at the University of Georgia (see also Kretzschmar, this volume) contains downloadable data sets associated with fieldwork-based traditional dialect atlases produced in the USA. The recent *Digital DARE* (*Dictionary of American Regional English*) project includes online maps representing a snapshot of lexical variation in American regional speech, as elicited by DARE fieldworkers from 1965 to 1970. The online *Atlas of Dialect Topography* (Chambers 2004) displays contemporary regional patterns for a small set of (largely) pronunciation and lexical features

¹ See: *BBC Voices*. Word Maps (<http://www.bbc.co.uk/voices/results/wordmap>), *Australian Word Map* (<https://www.macquariedictionary.com.au/resources/word/map>), *Harvard Dialect Survey* (<http://dialect.redlog.net>), *Cambridge Online Survey of World Englishes* (http://www.tekstlab.uio.no/cambridge_survey).

² The BBC Word Maps can be viewed at <http://www.bbc.co.uk/voices/results/wordmap>. Katz' maps are available at <http://www4.ncsu.edu/~jakatz2/project-dialect.html>.

in eastern Canada and neighbouring American states, with data derived from self-reporting mail-in questionnaires rather than fieldwork. Labov et al.'s (2006) monumental phonological *Atlas of North American English*, based on telephone interviews, uses acoustic and statistical analyses to delineate major dialect areas in contemporary North American speech, relative to ongoing linguistic changes.

Novel computerized approaches have also figured prominently in the study and representation of regional variation in a range of languages other than English. To mention just a few, these include the pioneering efforts of Embleton et al. (2001, 2007) and Heap (2003), who have converted traditional regional dialect materials to online atlases—the first, for Romanian, and the second, for the speech varieties of the Iberian Peninsula. Contemporary regional speakers are represented in the *Syntactic Atlas of the Dutch Dialects* (SAND) project (Barbiers et al. 2007). Junker et al.'s *Algonquian Linguistic Atlas* (2005) has used open source software and Google mapping tools to produce an online, multimedia site targeting contemporary Canadian aboriginal students.

Linguists have also developed innovative computer techniques and approaches to create online regional dialect maps. Noteworthy among these is the Gabmap web application (see Nerbonne et al. 2011), designed to easily enable quantitative analysis and display of dialect data. Grieve et al. (2013) advocate an approach using commercial search engines (such as Google) to construct contemporary lexical maps based on a restricted set of regionally representative websites (in their case, online newspapers); they conclude that this quick approach to regional dialectology compares favourably with the traditional dialect maps derived from fieldwork-administered questionnaires.

This chapter presents a new online regional linguistic atlas, launched in late 2013: the *Dialect Atlas of Newfoundland and Labrador*. This atlas documents the traditional vernacular English spoken in one of Britain's oldest overseas colonies, which since 1949 has constituted Canada's youngest province. Within North America, the province's speech is unique in terms of its highly circumscribed founder varieties (originating primarily in southwest England and southeast Ireland), along with its considerable regional variation. Unlike many of the online regional dialect sites noted above, the focus of this atlas is not contemporary speech.

Rather, it is grounded in legacy recordings, both archival- and fieldwork-based, and represents the conservative varieties of older rural speakers of some three to five decades ago. It differs from traditional dialect atlases, however, by attempting to bridge the gap between publicly and academically oriented dialect-mapping endeavours.

In the spirit of Wolfram's principle of 'linguistic gratuity' (see Wolfram et al. 2008), our website engages with contemporary speakers in unique ways. Rather than simply presenting the phonetic, morphosyntactic and lexical features of speakers several generations removed from those of today, our atlas attempts to bring their speech to life. This it does through the incorporation of such components as illustrative sound files for phonetic variables, as well as interactive games designed to familiarize site visitors with the linguistic features of the traditional speech of Newfoundland and Labrador. But while our aims are to promote cultural heritage by engaging current generations, we have not done this at the expense of academic rigour: academic audiences will find all the information on linguistic features that they would expect from more traditional dialect atlases.

In Sects. 2 and 3 below, I outline the origins of the *Dialect Atlas of Newfoundland and Labrador* (referred to henceforth simply as the 'Atlas'), in its pre-digital and early digital phases. Sect. 4 shows briefly how the Atlas works, and details the development of the Atlas as a public resource, touching on the various themes of this volume, among them public outreach, social impact, sustainability and accountability of use.

2 The Dialect Atlas of Newfoundland and Labrador: Pre-digital Phases

The Atlas has been four decades in the making. Its foundations were laid in 1974, when Memorial University linguist Harold Paddock obtained the first of two Canada Council (now Social Sciences and Humanities Research Council of Canada) research grants to undertake a survey of regional linguistic differences in Newfoundland and Labrador. Paddock envisaged this as a purely academic undertaking. Its goal was an eventual print atlas documenting regional lexical, phonetic and grammatical

differences throughout the province, in the tradition of the Survey of English Dialects (Orton and Dieth 1962), along with such American dialect atlases as those of Kurath et al. (1939–1943), McDavid et al. (1979–) and Pederson et al. (1986–1993). Given such factors as its early settlement, small and extremely circumscribed European founder populations, relative geographic isolation, and highly conservative regional varieties (see for example Clarke 2010, 2013), Newfoundland and Labrador offers an ideal location for the investigation of regionally based variation.

2.1 Phonetic and Grammatical Components

In the first stage of the project (‘A Preliminary Dialect Mapping of the Island of Newfoundland’), Paddock’s focus was on ‘structural’—that is, phonetic and grammatical—features, rather than vocabulary. The rich repository of audio (reel-to-reel and cassette) recordings available in the Memorial University of Newfoundland Folklore and Language Archive (MUNFLA), most of them collected in the 1960s, obviated the need for the usual period of extensive fieldwork required by traditional dialect atlas projects to gather an adequate corpus of regional phonetic and grammatical features. Rather, the project relied on recordings of traditional speakers in 69 different communities (see Fig. 4.1), represented by 80 members in total.³ These represented the entire coastline of the island, along which virtually all Newfoundland communities are situated, given the area’s traditional economic dependence on the inshore cod fishery.

With very few exceptions (notably the cities of St John’s and Corner Brook), the settlements chosen were small and rural, with populations ranging from several dozen to under 3000. Speakers on the whole displayed the ‘NORM-like’ characteristics typical of respondents in traditional dialect surveys (though women constituted 25 per cent of the sample). Most would have had little schooling, having entered the fishery

³Most communities in this phase of Paddock’s survey were represented by a single speaker. The high degree of source dialect homogeneity within most small rural Newfoundland communities, however (see for example Clarke 2010: 10ff.), along with the conservative nature of rural Newfoundland speech, suggest that a fully representative sample of local speech features was obtained.

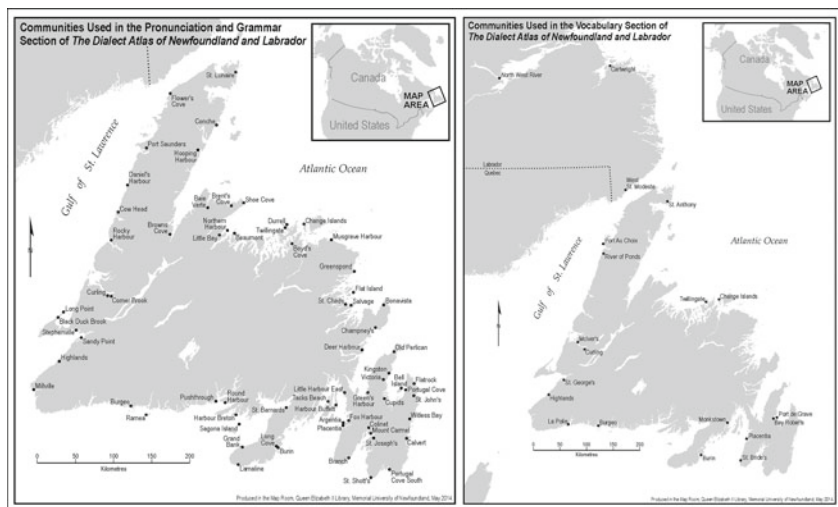


Fig. 4.1 *Left:* A map illustrating the communities used in the pronunciation and grammar section of the *Dialect Atlas of Newfoundland and Labrador*. *Right:* A map illustrating the communities used in the vocabulary section of the *Dialect Atlas of Newfoundland and Labrador* (Produced in the Map Room, Queen Elizabeth II Library, Memorial University of Newfoundland).

and/or participated in other subsistence and household-related activities to help support their families from a very young age. The oldest was born in 1871; all but 24 were born before 1900, and half of the post-1900 group have birth dates before 1910. These speakers are then comparable in age to those represented in the Survey of English Dialects, mentioned above, though on average perhaps a decade younger.

The recordings selected represent informal interviews, typically focusing on aspects of traditional folk life, including folk customs and beliefs; a small portion also involved the performance of traditional folk tales and folk songs.⁴ Given the social and regional profiles of the respondents, their speech is on the whole highly conservative and non-standard,

⁴One obvious shortcoming of the dependence on archival recordings rather than field interviews is that, as in this case, few appropriate recordings may be available for given areas, and those that are may not illustrate all the linguistic features under investigation. For example, in the case of one tiny community representing the French-settled area of the island's west coast, the sole recording available was of a well-known fiddle-player, and the recording consisted primarily of his music rather than speech.

preserving many of the regional features that would have characterized their southwest British and southeast Irish ancestors. Some of the flavour of this speech can best be captured by way of example. The following set of instructions on how to launch a boat come from a recording of one of the sample members, a fisherman from the northeast coast of the island, born in 1882 (MUNFLA C-299, 66–25). This short excerpt is presented for the most part in standard transcription, with the obvious exception of vowel elision and the notation of *-ing* as *in'*. These impart a better sense of the speaker's highly vernacular pronunciation.

You blocks up now with square stuff ... and ah you gets your long 42 one and puts 'in (*it*) like you wants 'in 'cross those gulches, and however the land is, you got the under ones now, to commence on, 'cross. Well you puts—you trigs up (*make trim*) your 42, your slipper I'll call 'in—no, he (*it*) isn't the slipper, he's the lanch (*launch*) way, the proper lanch way. You got—you trigs up he too and commence to block 'in, block 'in in. You lose that right over, you'm vastenin' (*you're fastening*) your blocks you know, 'cardin' as (*according as = while*) you comes up, vasten 'em on 'cause 'tis, 'tis nearly all solid ... And you blocks up all the length now to go out in the water under the schooner. When you gets that all done, you—gets it to your likin' and vastened and boltied (*bolted*) on...

Not only does this speaker make considerable use of non-standard present-tense verbal *-s* (as in *you blocks, you gets*), he also exhibits many conservative morphosyntactic features which attest to his southwest English ancestry. These include grammatical gender for inanimate nouns (*no he isn't the slipper, he's the lanch way*); the use of 'in (presumably from OE *hine*) rather than (*him* as a 3sg object pronoun; pronoun exchange, or use of subject-like pronominals in stressed object position (*you trigs up he too*); and *am* rather than *are* in persons other than the 1sg (*you'm vastenin'*). A noticeable southwest English feature of pronunciation is his voicing of initial fricatives, as in *fasten* pronounced *vasten*, with initial [v]. In addition, he employs many traditional and regional lexical items (*commence* for *begin*, *trig up* for *make trim* or *smart*), along with specialized maritime vocabulary.

Paddock's team searched the MUNFLA recordings for 46 features of traditional Newfoundland speech (21 phonetic, 25 grammatical),

Post-vocalic -l

	clear/dark. dark central/irish/ada.	Croke M1880.	hill, call, fall, vessel, girl [ɫ], top, oil [ɫ]	
1	WB, N	Spring Hill M1919	sell, help [ɫ], fall [ɫ], sell, tail [ɫ] all [ɫ]	
	dark central/irish/ada.	St. Ann's M1910.	twelve [ɫ, ɫ], small, well, people [ɫ]	
2	UBS	Bray's Cove F1881	well, sell, fall, oil [ɫ], old [ɫ]	
	dark	Breat's Cove F1903	all, call, vessel, fall [ɫ], old [ɫ]	
	dark	Breat's Cove F1903	well, sell, fall, oil [ɫ], old [ɫ]	
3	GB	St. Ann's M1883.	well, sell, fall, oil [ɫ], old [ɫ]	
	dark, central/irish	Northon In M187	well, sell, fall, oil [ɫ], old [ɫ]	
4	Low	Bayl's Cove M1920	dollar, well, fall, oil [ɫ]	
	dark			
5	TW	Tur M1891	well, you'll [ɫ]	
	dark	Myragrant Nr. M1886.	call, saddle [ɫ], roll [ɫ]	
6	TRG	St. Ann's M1886.	joily, all, sell, fall [ɫ]	
	dark	St. Ann's M1878	well, call, pull [ɫ]	
7	BN	Greenpod M1886	mail [ɫ]	
	dark, central/irish	Bonavista M1874	bill, foot, string [ɫ] spell [ɫ]; work, minute [ɫ]	
8	BS	Salvage F1871	all, call [ɫ], old school [ɫ], else [ɫ]?	
	dark, clear/ada			
9	TN	Champanys M1882	call, hill [ɫ], old [ɫ]	
	dark			
10	TS	Greay's Nr. F1882.	sell [ɫ], old [ɫ], red, call [ɫ], all [ɫ]	
	dark/clear	Old Pavilion M1896.	call, hill, mail, foot [ɫ], well, month [ɫ], suit [ɫ]	
11	BU	Kingston F1898.	old [ɫ], hill, school, call, people, girl [ɫ]	
	dark/clear			
12	Car	Victoria M1901.	all, well, hill, real [ɫ], old [ɫ], sell, help [ɫ]	
	clear/dark central/ada			
13	HG, BG	Cupido M1880	old, dull [ɫ], people, tall, real, mail, call [ɫ]	
	clear/dark			

Fig. 4.2 Secondary worksheet (partial) for postvocalic /l/, Paddock survey

including most of those noted above. Of these, 21 (12 phonetic, 9 grammatical) appeared to display obvious regional patterning, largely reflecting founder dialect input from the British Isles and Ireland. These were selected as potential candidates for mapping. Data were extracted manually directly from the cassette and reel-to-reel recordings, and were stored as handwritten 'worksheets': both 'primary' (per speaker) and 'secondary' (aggregated by linguistic variable). The latter is illustrated by Fig. 4.2, which represents a partial worksheet for the phonetic feature of postvocalic /l/.⁵ Only two maps were to appear in print (see Paddock 1982). One of these represents the regional distribution of allophones of postvocalic /l/; the second is the grammatical feature of pronoun exchange, as described above.

⁵ In Fig. 4.2, the leftmost column following the numbers represents geographical distribution by electoral district (for example, #8, BS, represents Bonavista South). The next column summarizes general findings for that area, in terms of Irish-like 'clear' or English-like 'dark' variants of /l/. This is followed by information on speakers' community, sex and birthdate (for example, Salvage F 1871). The final column contains tokens illustrating the variable, along with an IPA transcription of the variant of postvocalic /l/ that each contains.

2.2 Lexical Component

In 1981, seven years after the initiation of the project, Paddock and his team began a second phase, focusing uniquely on lexicon. To this end, an extensive dialect questionnaire was devised. Grounded in the questionnaires of Kurath et al. (1939) and Orton and Dieth (1962), it was considerably adapted to the local situation, particularly by targeting terms relating to the fishery, the sea and marine life. Its over 500 questions were also designed to elicit everyday vocabulary in such general semantic areas as farming, animals, nature, the house/housekeeping, the human body and social activities. A small number of these questions aimed additionally at obtaining lexical items of phonetic or morphological interest.

Resources did not permit a full fieldwork-based survey of the 69 island communities analyzed in the project's initial phase. Rather, Paddock divided the island into the eight linguistic regions suggested by the structural survey, supplemented by two areas in the continental portion of the province, Labrador. In each of these ten regions, a representative pair of communities was selected: one smaller and more 'rural', the other somewhat larger and more 'urbanized', often serving as a commercial centre for its immediate region (see the right-hand map in Fig. 4.1 for the location of these 20 communities).⁶ Fieldworker Kathleen Manuel administered the questionnaire both on the island and in Labrador; she made no notes on the responses provided, but simply recorded the entire interview on cassette tape. In each community, she generally interviewed six older traditional speakers, three males and three females.⁷ They were selected so as to ensure representation of the range of occupational activities typically available in such communities, and hence the specialized vocabularies associated with each: the trades (usually carpentry), the sea (fishing, fish processing) and the land (logging, farming).

⁶ As the province's total population is only half a million, the largest of the 'urbanized' communities sampled contained under 6000 residents; most, however, were no more than half that size. Given the project's focus on the traditional speech of 'NORMs', none of the province's larger communities were included in the lexical portion of the Atlas.

⁷ The final sample consisted of 124 rather than 120 respondents, as a result of inclusion of extra speakers who had relocated from one of the rural coastal Labrador communities surveyed to the more 'urbanized' Labrador community (North West River) represented in the Atlas.

By manually extracting responses to the questionnaire directly from the cassette recordings, Paddock's team produced three preliminary maps (*dragonfly*, see Paddock 1984; *sap of fir trees*; and *needles of conifers*). Further progress was hindered, however, by a number of factors: the sheer volume of responses (well over 40,000, in over 200 hours of recorded interview); the lack of sufficient long-term funding for data extraction and analysis; and the complexities that would have been associated with computerized data manipulation in the early to mid 1980s, when data coding and entry involved punch cards and mainframe computers. As a result, Paddock's huge project was put on hold. It was to languish for almost 20 years.

3 From Archival Data to Online Atlas: The Initial Steps

By the turn of the twenty-first century, the enormous strides in digital technologies of the previous two decades offered possibilities that had not been envisaged when the project was initiated. Around 2003, a team of researchers associated with Memorial University's English Language Research Centre (the present author, a linguist; folklorist Philip Hiscock; and English language specialist Robert Hollett) took it upon themselves to transform Paddock's materials into an online format. To this end, the research team enlisted the aid of Memorial University geographer Alvin Simms to provide expertise in Geographical Information Systems (GIS) and online mapping. In light of the enormous interest in local speech on the part of residents of the province, along with the large Newfoundland and Labrador diaspora, we envisaged from the start a website of interest and appeal not only to academic researchers (in the manner of Embleton et al. 2001, 2007; Heap 2003), but also the public at large.

Our lack of experience led us to estimate that this process would require a commitment of perhaps four years. However, it was not until a decade later that the online Atlas was formally launched. The delay can be attributed to three primary factors: (i) our discovery that even the structural data that had been extracted required re-extraction and extensive rechecking; (ii) the development of the various components needed to transform a purely academic resource into a 'user-friendly' one,

particularly the provision of illustrative sound files for every phonetic token extracted; and (iii) the difficulty of finding sufficient funding for a project that focused primarily on the preservation of cultural heritage rather than on research as generally understood by the scientific community. All of these points will be touched on in the course of this chapter.

3.1 Digitization

As Sect. 2 indicates, the two atlas surveys undertaken in the 1970s and 1980s did not yield any digital data. The existing handwritten ‘structural’ summaries had been extracted directly from non-time-stamped archival analogue recordings. As to the lexical data, not even summary records existed; the three preliminary maps produced had drawn their data directly from the recorded interviews. In fact, the extensive lexical questionnaire itself existed only in handwritten form. As a result, we found that we were starting virtually from scratch. Consequently, among our very first steps was the digitization of the handwritten instruments associated with the original surveys, notably the summary phonetic and grammatical worksheets, along with the lexical questionnaire. In addition, the MUNFLA archive contained typescripts for many of the interviews which served as the source of the phonetic and grammatical data. As we did not have the financial resources to retranscribe the contents of the MUNFLA recordings, or even to redo the typescripts as computerized word-processed documents, we had them scanned and converted to searchable PDF format. This enabled us to locate potential tokens of each linguistic variable much more easily—despite the lack of consistent transcription protocols, and obvious errors and omissions, in the original typescripts.

The second aspect of our early digitization involved the recorded interviews on which the two surveys were based. All of the more than 300 cassettes and reel-to-reel tapes associated with both surveys were converted to WAV format by means of the digital audio editing software Sound Forge 8.0, a commercial program created by the company Sonic Foundry (acquired subsequently by Sony).⁸ As is typically the

⁸We are extremely grateful to Sonic Foundry for providing us a free copy of its earlier versions of this software for the Atlas project, and to Sony for enabling us access to its updated 8.0 version. As an excellent alternative, the digital editing program Audacity is available as free open source software.

case of archival recordings produced by non-linguists, few of the original MUNFLA recordings had been made under optimal conditions. Interviews typically took place in the interviewees' homes, contained background noise, and were often conducted by folklore students and local interviewers with little knowledge of ideal recording procedures. As a result, most required considerable audio editing and clean-up via the Sound Forge program.

3.2 Data Extraction

Once digital audio files were created for both components of the project, full-scale data extraction could get underway. For the lexical component, responses to the questionnaire were extracted by research assistants directly from the digitized sound files, and transferred to lexical coding sheets for each of the 124 respondents. As to the structural (phonetic and grammatical) components, the original worksheet summary data were rechecked in detail against the digital recordings, and the resulting data stored in computerized spreadsheet format for direct incorporation into the developing project databases. Since the original IPA transcriptions of phonetic tokens illustrating each regional pronunciation had been done by Paddock's research assistants three decades earlier, checking was obviously required. This was undertaken by the three principal researchers, who used the opportunity of revisiting the digitized sound files to extract and transcribe thousands of illustrative tokens. This process also enabled considerable expansion of the number of phonetic features originally envisaged, from 21 to 31; likewise, the number of grammatical features was increased from 25 to 27.

3.3 From Data to Online Display

The conversion of our phonetic, grammatical and lexical data into onscreen map displays required two essential components. The first, computerized relational database software, enables the data to be stored in a series of relations or linked 'tables'. Each table represents a particular type of information (linguistic data, speaker, community and so on), and

data are retrieved and managed via structured query language (SQL).⁹ The second component consists of GIS mapping software, which interfaces with databases to produce dynamic online displays.

In both cases, concerns of efficiency and cost-effectiveness led initially to our adoption of commercial software. We selected Microsoft Access (part of MS Office Suite 2003) as the project's database program, along with the GIS software *AspMAP 3.0*, from VDS Technologies. The latter uses ASP (Active Server Pages) scripting language to query both spatial and non-spatial data, and can display results in various online formats, including maps, graphs and tables. The obvious concerns of longevity and sustainability associated with ever-changing commercial products, however, were to lead to our eventual switch to open source software (see Sect. 4.5.1. below).

4 Developing the Atlas as a Public Resource

4.1 Funding

The online Atlas experienced a fairly slow start. The reason for this was largely pragmatic: the availability of potential funding sources. With a primary focus on heritage preservation and knowledge dissemination, as opposed to theoretical advancement via new data collection and analysis, the project was not an obvious fit with the research programs offered at the turn of the twenty-first century by Canada's national granting agency, the Social Science and Humanities Research Council of Canada (SSHRC). Nor did the Atlas project prove an ideal fit with Canadian funding programs designed for the preservation of cultural heritage. At the provincial (Newfoundland and Labrador) government level, such funding is almost entirely dedicated to existing heritage and community groups. At the national level, funding offered by the Department of Canadian Heritage tends to target minority language groups; the funding programs of Library and Archives Canada and the Canadian

⁹ Reasons for the choice of a relational database for data storage are well articulated by Barbiere et al. (2007: 69ff.), who describe its application to their *Syntactic Atlas of the Dutch Dialects*.

Council of Archives have been undergoing severe cuts under the current federal government. (In fact, the latter's National Archive Development Program was totally eliminated in early 2012.) In short, the online Atlas project fell through the funding cracks. As a result, the project made steady but slow progress for the first half dozen years of its existence thanks to small research grants from Memorial University sources (the Institute of Social and Economic Research and the J.R. Smallwood Foundation), coupled with various university- and province-sponsored programs that provided part-time undergraduate and graduate student research assistance.

In 2010, however, we became aware of two large-scale national funding opportunities whose primary goals tallied extremely well with our aims of public outreach. The first, the Canada Interactive Fund, was offered by the federal government's Department of Canadian Heritage; the second, the Public Outreach Dissemination grant program, by SSHRC.¹⁰ A successful application to the latter provided us with the funding necessary to engage the experienced computer programmers and web designers needed to transform the existing website into one with considerably greater visual appeal and user-friendliness, and to develop it as an interactive resource. We found the ideal technical expertise in two Memorial University units: CCWebworks, the Web programming branch of Computing and Communications, and DELTS (Distance Education, Learning and Teaching Support). The latter's experience with online course design and delivery for both the university and the broader Newfoundland community proved particularly beneficial with respect to achieving our outreach goals of targeting not just the public at large, but, more specifically, cultural and heritage groups, along with school and university students.

Since the essential shape of the online website emerged during the Atlas's university-funded years, the final version of the site has built on this basic design. A brief overview of how the Atlas functions is provided in Sect. 4.2 immediately below. The ensuing Sects. (4.3–4.5) outline the

¹⁰We note with interest that, despite the increasing importance of public engagement to university-based research and university mission statements, both of these programs have since been discontinued, as has the Department of Canadian Heritage's Canadian Culture Online Program. SSHRC has, however, recently developed a 'Connection' program, designed to promote knowledge mobilization by developing networks and tools that can be utilized by non-academic audiences.

various ways in which, thanks to our SSHRC Public Outreach grant, the online Atlas has been developed into a heritage website with broad public appeal, as well as a teaching and learning tool.

4.2 How the Atlas Works: A Brief Introduction

Upon entering the Atlas portion of the project website at www.dialectatlas.mun.ca, visitors can select which of three types of regional language variation they would like to explore: pronunciation, grammar or vocabulary. In each case, they have access to a pull-down menu. The pronunciation and grammar menus provide a listing of the individual linguistic features investigated by the Atlas, and available for map display (an illustration is provided in Sect. 4.3.1). In the case of vocabulary, the menu lists the 16 semantic areas into which we reorganized the original questionnaire to make it more user-friendly; subsequent menus display subareas within each of these, followed by the full list of questions associated with each subarea. Table 4.1 provides an illustration using the semantic area ‘House and household’. Clicking on this option yields four semantic subareas; selecting each of these, in turn, brings up the list of questions it contains. Table 4.1 illustrates with one or two questions from each subarea.

If a site visitor selects the second ‘Food-related’ question above, and then clicks on the response *barm*, the result is the map that appears in

Table 4.1 Extract from the ‘House and household’ section of the lexical questionnaire

Semantic area	Subarea	Sample questions
House and household	Food-related	What do you call bread dough which has been fried up like a pancake? What was used in the old days to raise bread before baking?
	House-related	In rainy or damp weather the wood of a door will WHAT? (get bigger)
	Measures of food	In the old days, how would butter be measured?
	Seasonal living	What do you call a tin can used outdoors on an open fire to boil water in?

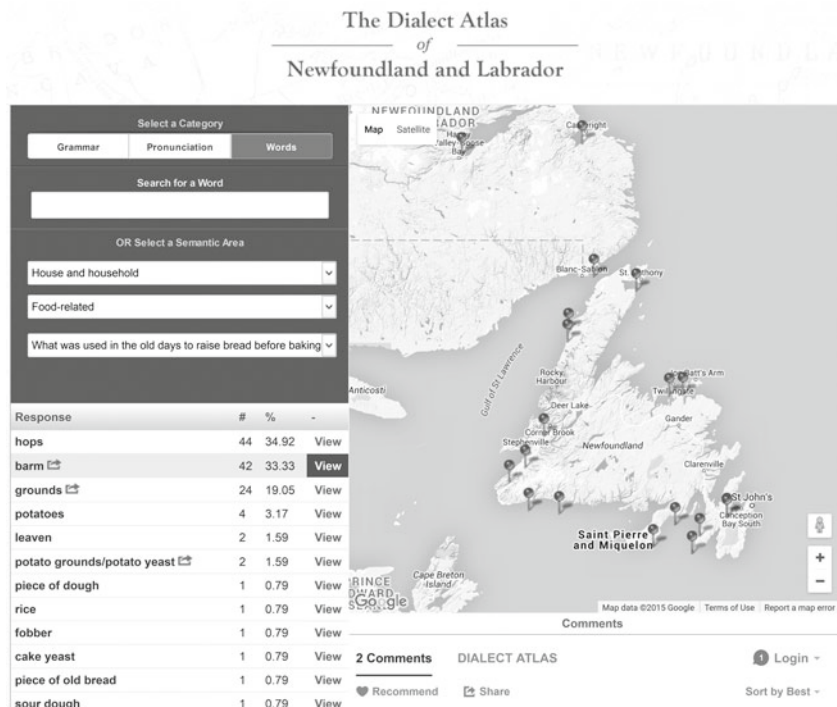


Fig. 4.3 Regional distribution of the term *barm* ('old-fashioned yeast') in Newfoundland and Labrador

Fig. 4.3. It indicates that this term was widely known throughout the province in the early 1980s, as it was returned in 19 of the 20 communities investigated. Clicking on the pin representing each community yields data on the gender and occupational breakdown of respondents, as seen in Fig. 4.4 for the term *piper*—a response to the 'Seasonal living' question in Table 4.1—which emerged for example in the island community, Port de Grave. The pie charts at the bottom left of Fig. 4.4 indicate the overall gender and occupational breakdown of the 79 respondents whose answers have been recorded for this question.¹¹

¹¹As suggested by the low number of responses to this question, not all lexical questions were answered by all interviewees. Further work would be required to fill these gaps. Since, however, more than 30 years have elapsed since data collection, most of the original interviewees are now deceased. While in such cases interviewing the current generation of community 'NORMS' should

The screenshot displays a web application interface for a dialect atlas. On the left, there are navigation options: 'Select a Category' (Grammar, Pronunciation, Words), a search bar, and 'OR Select a Semantic Area' (House and household, Seasonal living, What do you call the tin kettle with a big bottom used outdoor). The main content is a table of responses for the term 'piper (kettle)/tin piper'. Below the table are two pie charts: 'Gender Distribution for Question' and 'Occupation Distribution for Question'. On the right, a map of Port de Grave is shown with a pop-up window displaying the 'Gender Distribution' and 'Occupation Distribution' for the term. The bottom section shows a comments area with '0 Comments' and a 'DIALECT ATLAS' header.

Response	#	%	-
kettle (camp/country/woods/tin etc)	36	45.57	View
piper (kettle)/tin piper	23	29.11	View
can/tin can/gallon can	4	5.06	View
slut	3	3.8	View
flat-ass (kettle)	2	2.53	View
hustler	2	2.53	View
tin billy	1	1.27	View
wesleyan	1	1.27	View
skeleton	1	1.27	View
hurry-up	1	1.27	View
pompey	1	1.27	View
quick	1	1.27	View
urn	1	1.27	View
glamour	1	1.27	View
bogie	1	1.27	View

Gender Distribution for Question

Gender	Count	Percentage
F	46	58.2%
M	33	41.8%

Occupation Distribution for Question

Occupation	Count	Percentage
Trades	29	36.7%
Sea	23	29.1%
Land	27	34.2%

Port de Grave
Response: piper (kettle)/tin piper

Gender Distribution Total: 3

Gender	Count
Male	1
Female	1

Occupation Distribution Total: 2

Occupation	Count
Land	1
Trade	1
Sea	0

Fig. 4.4 Response breakdown for the term *piper* ('tin kettle') in the community of Port de Grave

provide further insight, so too would consultation of the huge card collection associated with the *Dictionary of Newfoundland English* (Story et al. 1982/1990), which is currently undergoing digitization for eventual online access (see http://www.mun.ca/elrc/projects/digitization_project.php).

4.3 Improving the ‘User-Friendliness’ of the Site

The Atlas’s online format obviously offers a wealth of opportunities to disseminate legacy linguistic content in ways which would not be possible in a traditional print atlas, and which appeal to a general as well as to an academic audience. For example, since many site visitors from Newfoundland and Labrador are more interested in the speech of their own communities and regions than in that of the province as a whole, we have recently added a ‘Search by community’ option that enables users to view only the results for the communities they select. Likewise, if Atlas users have a particular local word in mind, they can bypass the pull-down menus described above and opt for the ‘Search for a word’ function, which displays map results for any word that they input directly if it occurs in our lexical databases. In the lexical section as well, since many of the regional items returned by queries may not be familiar to present-day speakers of Standard English, we have included links to word definitions in the online *Dictionary of Newfoundland English* (Story et al. 1982/1990) whenever these appear in the Dictionary. The symbol to the right of the response *piper* in Fig. 4.4, for example, yields the DNE entry for the word, part of which is reproduced in (1).

(1) An extract from the linked definition for *piper* in the *Dictionary of Newfoundland English* online (<http://www.heritage.nf.ca/dictionary/#3375>)

piper n *ADD* ~ Nfld (1921). Locally made galvanized tin kettle with bevelled sides, flat bottom and a long narrow spout or ‘bib’ emitting a piping sound when it boils; BIBBY, QUICK, SLUT. [c1845] 1927 DOYLE (ed) 27 “The John Martin”: He took the kettle by the hangers, and he threw it on the ice; / I never felt so scalded since the day that I was born, / When I saw my little piper and it floating off astern. [...]

It is the more technical components of the Atlas, the phonetic and grammatical elements, which particularly required extensive development for a broad audience. The following subsections outline some of the chief features that have been incorporated with the general public in mind.

4.3.1 Sound Clips to Illustrate Pronunciation Features

One dimension of the Atlas which non-academic users have reacted to very favourably is that it does not simply describe regional phonetic (*Pronunciation*) features across island communities. Rather, it also provides dynamic audio file illustrations for every phonetic example, thereby making the Atlas come alive via the actual voices of speakers contemporaneous with listeners' great (or great-great, or even great-great-great)-grandparents' generations. To this end, some 4000 illustrative sound clips, each approximately two to five seconds in length, were extracted from the digitized sound (WAV) files (see Sect. 3.1). To enhance Web access speed, these were converted to the smaller, compressed MP3 format. When users select a phonetic feature and bring up the map documenting its regional distribution, clicking on any community pin results in a display of the actual tokens that exemplify that feature, as uttered by the speaker(s) from the community selected. This is illustrated here for the feature of /ay/ rounding in the south coast community of Round Harbour. Fig. 4.5 shows two responses from this settlement, *died* and *nine*, each of which has a rounded schwa nucleus, as the IPA symbols indicate. Clicking on the icon beside the token *nine* brings up not only the sound clip in which this pronunciation can be heard, but also a transcription of the content of the clip ('Oh not very large. Nine families, sir'; see the inset panel in Fig. 4.5). Without transcriptions, the sound files would in many cases be quite difficult to understand for those without a good knowledge of traditional Newfoundland speech.

Though the sound dimension of the project is among the most interesting for non-academic users, it required by far the greatest investment of human resources, involving as it did thousands of hours of research time. Given the highly localized nature of the speech, and the often far from optimal quality of the original recordings, the sound files had to be checked and rechecked for content not only by student assistants, but also by the three principal researchers. Difficult clips were flagged, and often engendered much debate as to content. Moreover, clips frequently underwent a series of cleanups via the Sound Forge audio editing program.

The screenshot shows the 'The Dialect Atlas of Newfoundland and Labrador' website. The main interface has a 'Select a Category' dropdown with 'Grammar', 'Pronunciation', and 'Words' options. Under 'Pronunciation', there is a 'Select a pronunciation feature' dropdown set to 'Vowel contrasts in TOY vs. TIE words'. Below this, there are two buttons: 'See where words like TOY and POINT can sound like TIE and PINT.' and 'See where words like TIE or PINT can sound like TOY and POINT.'. A map of Newfoundland and Labrador is visible in the background. Two pop-up windows are shown: one for 'Round Harbour' with 2 responses, displaying a table of words and pronunciations, and another for 'Round Harbour' with 2 responses, displaying a transcript and an audio player.

Word	Pronunciation
died	[ɹɪ]
nine	[ɹɪ]

Transcript: 'Oh, not very large. Nine families, sir'
(Response: nine)
Play the audio file

Fig. 4.5 TIE rounding, with illustrative tokens from Round Harbour. Inset: Round Harbour sound clip and transcription for the token *nine*

4.3.2 ‘Translating’ Technical Vocabulary for a General Audience

As noted earlier, the Atlas began as a purely academic enterprise, and through its early years as a Web resource it maintained much of its academic feel. The entry of Memorial University’s DELTS unit as the site’s principal technical developer in 2011, however, led to a substantial revamp. This involved not only major changes in the overall look of the site, but also in the treatment of the technical linguistic vocabulary required for representation and discussion of phonetic and grammatical features. Where possible, such vocabulary was *translated* into more generally comprehensible terms. Where this proved impossible, every

effort was made to provide explanations that could be understood even at the intermediate school level, since among our target groups are 13- to 14-year-old students who we hope will make considerable use of the Atlas in their Newfoundland and Labrador Social Studies courses.

One obvious need for change lay with the descriptors of some of the pronunciation and grammatical features examined in the Atlas, along with the wording of database queries used to produce map displays. Many of the Atlas's initial variable names, though perfectly comprehensible to linguists, proved well beyond the capacity of an expanded audience. As a result, with the input of DELTs' curriculum specialists, we have opted to use such well-known and more comprehensible descriptors as 'Long A' instead of /eɪ/, 'Short I' instead of /ɪ/, and to describe vowel contrasts and mergers in such terms as 'Vowel contrasts in COT and CAUGHT words' and 'Vowels before R: FEAR vs. FARE words'. Table 4.2 illustrates the type of wording changes we have made to the queries, using two pronunciation and two grammatical examples.

In cases where technical vocabulary is unavoidable, we have attempted to provide concise explanations that are within reach of most. An information icon appears onscreen for every phonetic and grammatical feature investigated in the Atlas. When this is clicked, it brings up an information box of the type illustrated in Fig. 4.6 for the 'Long A' vowel variable.

The first paragraph of the box constitutes a general explanation of the feature in question. Subsequent paragraphs (only the first of which is illustrated here for 'Long A') typically provide further historical information on the feature, and outline the regional distribution of its variants in Newfoundland, as indicated by our online maps. Definitions of any technical terms—such as 'monophthongs' and 'diphthongs' in the second paragraph above, displayed in blue font onscreen—can be brought up by rolling the mouse over the term: 'monophthongs', for example, are defined as '*single* vowels, in which there is no change in quality during production (e.g. the *-i-* of *pit* or the *-u-* of *but*)'.

In our choice of wording, we have striven to find a balance between a linguistic and a general audience. We feel that we have not sacrificed linguistic rigour in the process: linguists should find all the information they would typically encounter in a print dialect atlas, including IPA phonetic

Table 4.2 Examples of wording changes in pronunciation and grammatical queries

Variable name	Initial query	Revised query
Short I vs. Short E vowel: TIN vs. TEN	In which communities do our data indicate that words in the standard English TEN class may be pronounced like standard English TIN?	See where words like TEN, BED or PET can sound like TIN, BID or PIT.
Long A vowel (MAID vs. MADE words)	In which communities do our data indicate that a distinction continues to be made between MADE and MAID types of words, through use of 'steady' or non-upglided pronunciations in the MADE type, but diphthongized/upglided pronunciations in the MAID type?	See where words like MADE or MADE can be pronounced either with a lengthened 'long A' sound (as in MAID pronounced 'ma-a-d'), or else in two syllables (as in MADE pronounced 'ma-yud').
2nd person pronoun forms: YOU, YOUS, YE, (D)EE	In which communities do our data indicate that the form YE occurs as a plural form of YOU?	See where YE can be used to speak to more than one person.
Forms of -SELF pronouns	In which communities do our data indicate that reflexive pronouns may be based on possessives: HISSELF, THEIRSELF/VES, along with MESELF?	See where MESELF can be used instead of MYSELF, HISSELF instead of HIMSELF, and THEIRSELF/THEIRSELVES instead of THEMSELVES.

transcription of all tokens of phonetic variables elicited.¹² In addition, researchers will, upon application, be able to access all our databases and sound files (see Sect. 4.5.2 below).

¹²The insertion of complex IPA symbols which display correctly across all platforms required some ingenuity on the part of our CCWebworks programmer. The symbols themselves derive from the following site, which provides an HTML equivalent for each: <http://symbolcodes.tlt.psu.edu/bylanguage/ipachart.html>. With our programmer, we developed a *translation table*, in which each individual phonetic symbol, including diacritics, was initially represented by a unique number which could then be converted to proper HTML code. (Complex symbols involving one or more diacritics were represented by the corresponding sequence of numbers.)

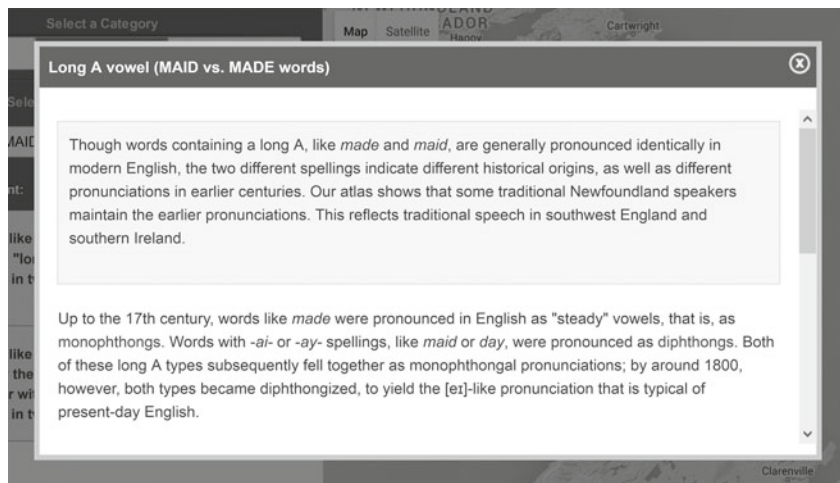


Fig. 4.6 Information box for the 'Long A vowel' variable

4.3.3 Interactivity: Games, Contributions/Comments, and Social Media Presence

Since we wished to build a Web 2.0 site that would actively engage users, rather than making them simply passive viewers of legacy content, one important feature of the online Atlas is its interactivity. Initially, we were hoping to include the option for site visitors to directly input data from their own communities, and thereby build up a section of the website consisting of dynamic maps. Unfortunately, this proved beyond our present resources, in terms of both funding and available personnel.¹³

Without doubt the most popular feature of the Atlas website is its interactive 'Activities' section. This section was developed by the DELTS technical team, using Flash and ActionScript. It consists of three games, each tied to one of the components of the Atlas: Words, Pronunciation and Grammar. This 'learning via play' dimension of the website, which

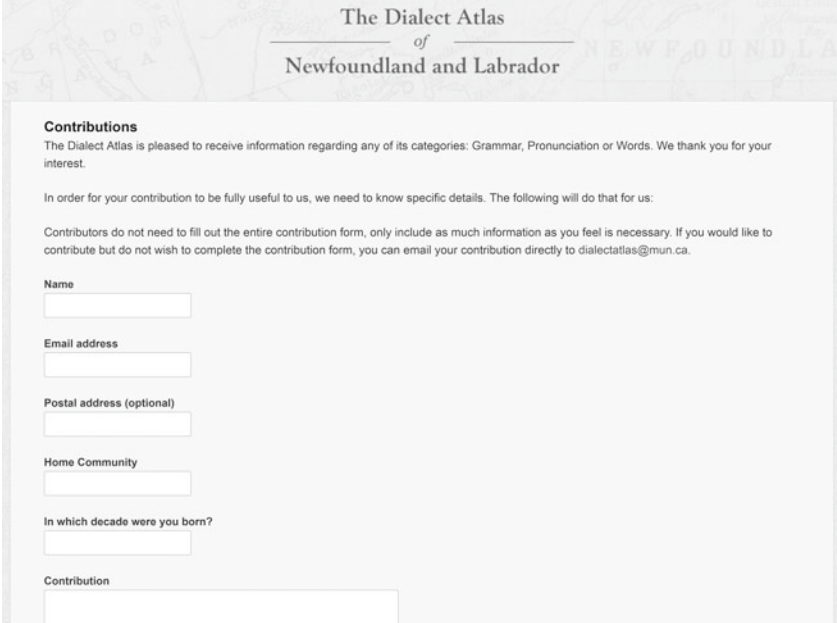
¹³Note that while the now inactive *BBC Voices Word Maps*, along with the *Australian Word Map*, did incorporate a dynamic dimension, this may prove difficult to manage. For example, the original plans for dynamic map updating of the *Cambridge Online Survey of Word Englishes* (Vaux and Jøhndal 2007) had to be put on hold as a result of the unusually high traffic generated by the site.



Fig. 4.7 Interactive games from the online Atlas

targets younger users in particular, aims to increase site visitors' familiarity with traditional Newfoundland and Labrador speech, as well as provide an enjoyable introduction to legacy linguistic material. The games associated with the Words and Pronunciation components of the site ('Word Jiggin' and 'Say What?', respectively; see Fig. 4.7) are timed activities. In the first, players drag and drop letters in an attempt to unscramble as many local lexical items as possible from the Atlas, with the assistance of a question hint (thus a question on the name for a particularly bad smell has the correct answer *fouisty*). In the second, users hear a succession of audio clips of Atlas speakers and try and match each with the correct transcription of its contents, from the set provided. The third game, 'Off the Top of My Head', draws on non-standard segments found in the Grammar section of the Atlas (examples include *after they was drowned; when I'm done of ye; I've a-been here; so I built a cellar for he; they stopped right sudden; it bees tough*). This game constitutes a creative writing activity in which users select from among these segments to compose (often humorous) stories, poems and songs. The results can be shared via the built-in copy feature.

With a view to increasing interactivity through building a community of online users, the Atlas site has been designed so that visitors can comment on any page. User comments (which are vetted by the site manager before being posted) also have the potential of yielding information



The Dialect Atlas
of
Newfoundland and Labrador

Contributions
The Dialect Atlas is pleased to receive information regarding any of its categories: Grammar, Pronunciation or Words. We thank you for your interest.

In order for your contribution to be fully useful to us, we need to know specific details. The following will do that for us:

Contributors do not need to fill out the entire contribution form, only include as much information as you feel is necessary. If you would like to contribute but do not wish to complete the contribution form, you can email your contribution directly to dialectatlas@mun.ca.

Name

Email address

Postal address (optional)

Home Community

In which decade were you born?

Contribution

Fig. 4.8 Partial view of a user contribution form

on communities not tapped in the surveys, and provide information on ongoing linguistic change. In addition, the site contains a contribution form (part of which is reproduced in Fig. 4.8), that encourages visitors to send in the local terms that they know and use, thereby enlarging our general body of knowledge on contemporary regional Newfoundland and Labrador vocabulary. While to date our contribution form has proved moderately successful, we are receiving fewer comments than we had hoped for. We are currently exploring ways of garnering more input through social media, since the Atlas has its own Facebook page and Twitter account. Among ideas under investigation is a Facebook ‘Atlas Word of the Week’ feature, which would list the regional variants that the Atlas has collected for selected concepts, and invite further contributions. In addition, the blog maintained by Memorial University’s English Language Research Centre—Twig, at twignl.wordpress.com—offers a means of reaching other potential Atlas contributors.

As an additional means of bringing the Atlas to the attention of the general public, DELTS has produced a promotional video, which they have posted on YouTube as well as on the ‘About’ page of the Atlas website.¹⁴

4.3.4 Access via Mobile Devices

The development period of the Atlas has coincided with a dramatic change in modes of online access, through the shift away from personal computers to handheld mobile wireless devices, smartphones in particular. If the Atlas is to appeal to a broad cross-section of the public, including students, it is imperative that it be accessible on the devices that they regularly use. As a result, the Atlas has been adapted to be fully accessible on wireless platforms. Since the original design was intended for desktop and tablet-sized devices, the site needed to be reformatted to display properly on phone-sized devices. This was accomplished through rewriting of parts of the HTML code and stylesheets so as to make the site *responsive*, that is, to adjust automatically how elements are displayed to accommodate smaller screens. The sole exception, however, is the interactive games component, which does not display on smartphones. This results from the use of Flash—at the time, the web standard in animation and multimedia tools—in the production of these games, since this was a platform in which our developers, DELTS, had a great deal of in-house experience. Unfortunately, Flash is not supported on all smartphones. In order to keep costs down, we decided not to redevelop the games in a different format for viewing on a smaller screen.

4.4 Social Impact

Over the past 70 or so years, Newfoundland and Labrador has experienced massive social and economic change. It has gone from a largely rural environment, in which the inshore cod fishery was the economic backbone of the thousands of tiny *outport* communities dotting the

¹⁴For the YouTube video, see <https://www.youtube.com/watch?v=10MSgbnYGjk>.

coast, to a more urbanized oil-driven economy, in which out-migration to larger regional centres, as well as elsewhere in Canada, has resulted in the loss of many coastal settlements. There is a widespread perception in the province that socioeconomic change is reflected in language change, and that local features are being increasingly edged out by more standard, supralocal ones. The distinctiveness of the province's regional dialects, along with the strong sense of place shared by most Newfoundlanders and Labradorians, means that there is intense interest in local speech. The province was the first in Canada to have its own dictionary (Story et al. 1982); its characteristic speech patterns are frequent topics in popular Web postings, as any Google search will reveal; and it has recently seen the development of a humorous downloadable smartphone *translator* entitled 'Whaddaya app' (a play on the local expression *Whaddaya at?*, with the general meaning of 'How's it going?'), designed for both iPhones and Android smartphones.

Consequently, one of our primary goals in developing the Atlas has been knowledge mobilization, through the transformation of legacy archival materials to an online format accessible to all. We see the Atlas as an important tool for the preservation of cultural heritage. Since its October 2013 launch, it has generated considerable interest and positive feedback from members of cultural and heritage associations, as well as from local and national media. One of our principal targets, however, is the education system. The Atlas is currently used as a resource in several courses taught at Memorial University. Recently, in conjunction with the province-wide English School District, as well as the Department of Education, we have developed a set of lesson plans for the Grade 8 Social Studies curriculum, which focuses on the province's changing history over the last two centuries. These materials have been widely circulated to teachers, and can be accessed from the Atlas website. The Atlas has also provided input to online materials designed by Memorial's ESL division for non-native-English-speaking professional newcomers to Newfoundland and Labrador, to help combat challenges to successful integration posed by the province's highly distinctive local dialects.

4.5 Long-term Sustainability

A major issue for regional dialect websites that target the general public is their long-term viability. Some however have proven to be largely short-term. For example, the *Australian Word Map*, a 2005 co-production of the Australian Broadcasting Corporation (ABC) and Macquarie Library, formerly online at the ABC website, was archived and retained purely for reference purposes. Its original URL quite recently became inaccessible, and the Word Map now exists simply as one of a number of resources on the Macquarie Dictionary online site.¹⁵ In similar fashion, the ‘Word Map’ section of *BBC Voices* is no longer dynamic, and has not been updated since October 2005. Good planning with respect to long-term curation and management (see Kendall 2013 for an excellent discussion) is required if websites are not to run the risk of a relatively short lifespan. The situation is complicated as site creators and managers move from one location to another, retire or die. By way of example, the *Harvard Dialect Survey*, created by Bert Vaux and Scott Golder, was hosted at Harvard University from 2002 to 2005. It then moved, with Vaux, to the University of Wisconsin, Madison, through which a non-dynamic version of the questions and maps can still be accessed. Vaux is now at the University of Cambridge. Fortunately, in this case, the Harvard Survey has been integrated into the ongoing *Cambridge Online Survey of Word Englishes*.

Issues of long-term viability have come to concern us more and more as we have developed the online Atlas, and we have taken a number of steps with a view to increasing sustainability. Among these, discussed below, is conversion to open source software, along with provision of access to our databases for interested researchers.

4.5.1 Use of Open Source Software

Initially based on commercial software (see Sect. 3.3 above), the project migrated in 2010 to widely used standards-based open source software, which should be viable for some time to come. Its data are now stored in

¹⁵ The original URL was <http://www.abc.net.au/wordmap>. The Macquarie Dictionary website URL is <http://www.macquariedictionary.com.au/resources/word/map>.

MySQL, among the most commonly used of all open source relational databases; its databases are queried and dynamic Web pages produced via PHP, a cross-platform-compatible open source scripting language that has become the standard in the production of dynamic Web pages. For its map displays, the project now draws on Google Maps, the norm for online mapping applications, which also offers the advantage of being easily customizable. All three of these choices ensure not only reliability but also flexibility, in that they run on a wide range of platforms, including Windows, Mac, Linux and UNIX. They also run in multiple browsers (Internet Explorer, Firefox, Chrome, Safari, and so on), as well as on portable handheld devices, which are becoming increasingly the norm for Web access.

4.5.2 Provision of Databases and Sound Files

With a view to ensuring that our data are shared, a ‘For researchers’ link on the Atlas website will allow password-protected access to our databases and sound files.¹⁶ The Atlas databases will be available in CSV (comma-separated-values) format, which can be read in a range of spreadsheet and database programs (as used, for example, on the *Linguistic Atlas Project* website at the University of Georgia). Our sound clips, in WAV and MP3 formats, will be available as zip files, subject to the restrictions noted in the following section. We also hope to make accessible the complete digitized interviews from which the lexical portion of the Atlas data has been derived, since these constitute a potentially rich source of linguistic information, well beyond the purely lexical data extracted in the development of the Atlas.

4.5.3 Long-term Preservation, Storage and Access

Currently, the Atlas website is managed by its content creators, all of whom are associated with the English Language Research Centre (ELRC) of Memorial University; the Atlas itself is stored on the servers

¹⁶As of January 2015, we are finalizing access forms based in part on those of similar websites such as the *Linguistic Atlas of the Iberian Peninsula* (see <http://westernlinguistics.ca/alpi>). As in the case of that site, use of our data will be restricted to research or teaching purposes, will require explicit acknowledgement of the source, and will be strictly not-for-profit.

of Memorial's division of Computing and Communications. However, issues of long-term management have yet to be worked out. These are of particular urgency as we the academic content creators of the Atlas retire or approach retirement, and as there is no guarantee that the ELRC itself will continue to survive, since the overall focus on local cultural heritage appears to be diminishing within the university's Faculty of Arts. Further, the ELRC's sister archive MUNFLA, with its folklore orientation, has not kept pace with best practices with respect to linguistic data digitization and management.

Kendall (2013) provides a good overview of issues relating to the preservation of, and access to, recorded legacy materials (see also Kendall and Wolfram, this volume). He points to the advantages to be gained through their deposit in a common archive, such as the *Sociolinguistic Archive and Analysis Project* (SLAAP), housed in the libraries of North Carolina State University. With the points raised by Kendall in mind, we plan to investigate options for long-term preservation of the online Atlas and the materials associated with it.

Hand in hand with preservation is the issue of accountable use of Atlas materials. As noted by the copyright link on every page, except for its sound files—which remain the property of Memorial University and the individual collectors—the online Atlas is subject to a Creative Commons non-commercial 3.0 license. This means that users are free to share (that is, 'copy and redistribute the material in any medium or format') and adapt ('remix, transform, and build upon the material'), except for commercial purposes, as long as they acknowledge the Atlas as the source. Such a restriction is impossible to enforce since, without encryption, all of the online materials, including sound files, can be downloaded by tech-savvy site visitors.¹⁷

¹⁷ A very obvious case of lack of proper attribution is documented by Zimmer (2013) with respect to the *Harvard Dialect Survey* (Vaux and Golder 2003). Its maps were reworked by Joshua Katz, a graduate student at North Carolina State University who did an internship at the *New York Times*; these 'went viral' and became the most popular piece of content published by that paper in 2013 (Graff 2014). While this is a situation where attribution was simply lost as the maps were popularized by both traditional and social media, it demonstrates the lack of content control inherent in much online publication.

5 Conclusion

In this chapter, I have reviewed the transformation of a set of legacy archival recordings into an interactive online dialect atlas documenting regional variation in the phonetic, grammatical and lexical features of traditional Newfoundland and Labrador speech. The rich linguistic heritage of the province, along with the intense interest of its residents in local speech varieties, has led us to attempt to create a website of as much appeal to the public at large as to the scholarly community. For present-day speakers—especially younger generations whose linguistic repertoires differ considerably from those represented in the Atlas—we hope to have provided a piece of living history, through access to the voices and speech patterns of generations past. From the scholarly perspective, our data should have much to offer, from potential insights into the conservative linguistic features of the province's chief source varieties in southwest England and southeast Ireland, to analysis of the areal and historical connections among regions of the province that are suggested by our linguistic databases.

With few models on which to base ourselves, our path to the interactive online Dialect Atlas has been a learning experience. We hope that our approaches, as shared in this chapter, may prove of some assistance to those interested in exploiting the rich resources that may also lie in their local speech archives.

References

Books and Articles

- Barbiers, Sjef, Leonie Cornips, and Jan-Pieter Kunst. 2007. The Syntactic Atlas of the Dutch Dialects (SAND): a corpus of elicited speech as an on-line dynamic atlas. In *Creating and Digitizing Language Corpora: Synchronic Databases*, vol 1, eds. Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, 54–90. Basingstoke: Palgrave Macmillan.
- Clarke, Sandra. 2010. *Newfoundland and Labrador English*. Edinburgh: Edinburgh University Press.

- . 2013. Adapting legacy regional language materials to an interactive online format: The Dialect Atlas of Newfoundland and Labrador English. In *Proceedings of Methods XIV. Papers from the Fourteenth International Conference on Methods in Dialectology, 2011*, eds. Alena Barysevich, Alexandra D'Arcy, and David Heap, 205–214. Frankfurt: Peter Lang.
- Grieve, Jack, Costanza Asnaghi, and Tom Ruetter. 2013. Site-restricted web searches for data collection in regional dialectology. *American Speech* 88(4): 413–440.
- Kendall, Tyler. 2013. Data preservation and access. In *Data Collection in Sociolinguistics*, eds. Christine Mallinson, Becky Childs, and Gerard Van Herk, 195–205. New York: Routledge.
- Kurath, Hans, with the collaboration of Marcus L. Hansen, Bernard Bloch and Julia Bloch. 1939. *Handbook of the Linguistic Geography of New England*. Providence, R.I.: Brown University.
- Kurath, Hans, with the collaboration of Miles L. Hanley, Bernard Bloch, Guy S. Lowman, Jr. and Marcus L. Hansen (ed.). 1939–43. *Linguistic Atlas of New England* (3 volumes). Providence, RI: Brown University.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English*. Berlin/New York: Mouton de Gruyter.
- McDavid, Raven I., Raymond K. O'Cain and George T. Dorrill (eds). 1979–. *Linguistic Atlas of the Middle and South Atlantic States* (2 volumes). Chicago: University of Chicago Press.
- Orton, Harold, and Eugen Dieth. 1962. *Survey of English Dialects (4 vols)*. Leeds: E.J. Arnold.
- Orton, Harold, Stewart Sanderson, and John Widdowson. 1978. *The Linguistic Atlas of England*. London: Croom Helm.
- Paddock, Harold. 1982. Newfoundland dialects of English. In *Languages in Newfoundland and Labrador*, 2nd edition, Harold Paddock (ed.), 71–89. St. John's, NL: Memorial University of Newfoundland. Also at http://www.dialectatlas.mun.ca/about/H%20Paddock_1982%20paper_final%20_Oct%2019-12.pdf
- . 1984. Mapping lexical variants in Newfoundland English. In *Papers of the Seventh Annual Meeting of the Atlantic Provinces Linguistic Association*, Helmut Zobl (ed.), 84–103. Moncton, NB: University of Moncton. Also at http://www.dialectatlas.mun.ca/about/H%20Paddock_1984%20paper_final_Oct%2019-12.pdf.
- Pederson, Lee, Susan L. McDaniel and Carol M. Adams (eds). 1986–1993. *Linguistic Atlas of the Gulf States* (7 volumes). Athens, Georgia: University of Georgia Press.

- Story, G. M., W. J. Kirwin and J. D. A. Widdowson (eds). 1982 [2nd ed. 1990]. *Dictionary of Newfoundland English*. Toronto: University of Toronto Press. Also at <http://www.heritage.nf.ca/dictionary>.
- Upton, Clive, Stewart Sanderson, and J.D.A. Widdowson. 1987. *Word Maps. A Dialect Atlas of England*. London: Croom Helm.
- Upton, Clive, and J.D.A. Widdowson. 1996. *An Atlas of English Dialects*. New York: Oxford University Press.
- Viereck, Wolfgang and Heinrich Ramisch (eds). 1991. *The Computer Developed Linguistic Atlas of England* (2 volumes). Tübingen: Niemeyer.
- (eds). 1997. *The Computer Developed Linguistic Atlas of England* (2 volumes). Tübingen: Niemeyer.
- Wolfram, Walt, Jeffrey Reaser, and Charlotte Vaughan. 2008. Operationalizing linguistic gratuity: from principle to practice. *Language and Linguistics Compass* 2(6): 1109–1134.

Websites, Software and Online Resources

- Australian Word Map*. 2005. ABC Online and Macquarie Library PTD Ltd. <https://www.macquariedictionary.com.au/resources/word/map> (accessed 1 June 2014); originally at <http://www.abc.net.au/wordmap>.
- BBC Voices Word Map*. 2005. <http://www.bbc.co.uk/voices/results/wordmap> (accessed 30 April 2014).
- Chambers J.K. 2004. *Atlas of Dialect Topography (On-line)*. http://dialect.topography.chass.utoronto.ca/dt_atlas.php (accessed 30 April 2014).
- Clarke, Sandra et al. 2013. *The Dialect Atlas of Newfoundland and Labrador*. <http://www.dialectatlas.mun.ca> (accessed 1 June 2014).
- Digital DARE (Dictionary of American Regional English)*. 2013. <http://www.daredictionary.com> (accessed 12 May 2013).
- Digital DARE (Dictionary of American Regional English)—DARE maps*. 2013. <http://www.daredictionary.com/page/maps/dare-maps> (accessed 12 May 2013).
- Embleton, Sheila M., Dorin Uritescu and Eric S. Wheeler. 2001. *Romanian Online Dialect Atlas (RODA)*. <http://www.yorku.ca/embleton/research/romanian> (accessed 5 May 2014).
- . 2007. *Romanian Online Dialect Atlas (RODA)*. <http://www.yorku.ca/embleton/research/romanian> (accessed 5 May 2014).
- Gabmap*. Information Science, Groningen and Meertens Institute, Amsterdam. <http://www.gabmap.nl> (accessed 30 May 2014).

- Graff, Ryan. 2014. Behind the dialect map interactive: how an intern created The New York Times' most popular piece of content in 2013. <http://knightlab.northwestern.edu/2014/01/20/behind-the-dialect-map-interactive-how-an-intern-created-the-new-york-times-most-popular-piece-of-content-in-2013> (accessed 12 May 2014).
- Heap, David. 2003. *The Linguistic Atlas of the Iberian Peninsula*. <http://western-linguistics.ca/alpi> (accessed 5 May 2014).
- Junker, Marie-Odile *et al.* 2005. *Algonquian Linguistic Atlas*. <http://www.atlasling.ca> (accessed 12 May 2014).
- Katz, Joshua. 2013. Beyond “Soda, Pop, or Coke”: Regional Dialect Variation in the Continental US (using data from the *Harvard Dialect Survey*). <http://www4.ncsu.edu/~jakatz2/project-dialect.html> (accessed 12 May 2014).
- Linguistic Atlas Project (LAP)*. University of Georgia. <http://www.lap.uga.edu>. (accessed 5 May 2014).
- Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg and Therese Leinonen. 2011. *Gabmap*—a web application for dialectology. <http://www.gabmap.nl/wp-content/uploads/2011/05/Gabmap-long-2011-jan-07-rev-mei.pdf> (accessed 16 May 2014).
- Sound Forge 8.0. 2006. Madison, Wisconsin: Sony Media Software. [computer program]
- Vaux, Bert and Scott Golder. 2003. *The Harvard Dialect Survey*. <http://dialect.redlog.net> (accessed 12 May 2014).
- Vaux, Bert and Marius L. Jøhndal. 2007. *The Cambridge Online Survey of World Englishes*. http://www.tekstlab.uio.no/cambridge_survey (accessed 12 May 2014).
- Zimmer, Ben. 2013. About those dialect maps making the rounds... <http://languagelog.ldc.upenn.edu/nll/?p=4676> (accessed 12 May 2014).

5

Engagement Through Data Management and Preservation: The North Carolina Language and Life Project and the Sociolinguistic Archive and Analysis Project

Tyler Kendall and Walt Wolfram

1 Introduction

In this chapter we consider the ways that enhanced data management and preservation practices improve our engagement processes and procedures. Data management and engagement are often considered separate professional enterprises with different goals and methods. Despite their disparate traditions, we argue that there is indeed an underlying intersection between these activities. We examine this possibility by demonstrating a model established for North Carolina, one of the most linguistically diverse states in the United States. For linguists, this natural linguistic diversity raises a number of empirical questions about the nature of language variation and change. At the same time, it also provides a captivating window for the

T. Kendall (✉)

University of Oregon, Eugene, OR 97403-1290, USA

W. Wolfram

North Carolina State University, Raleigh, NC 27695, USA

public into the intersection of language and society, with social and educational implications for students, the citizens of North Carolina, and the wider public. For more than two decades, the *North Carolina Language and Life Project* (NCLLP) has strived to align the scientific study of language variation and change with the need for promoting public knowledge about language variation (Wolfram et al. 2008).

Documenting the linguistic landscape of North Carolina—or of any place—is an arduous and time-consuming task. The formal research we conduct, however, is often obscured in scholarly journals and in library archives, usually inaccessible to people who do not have graduate degrees in linguistics. But socially responsible professionals have now been challenged by positions that include the *principle of debt incurred* (Labov 1982) and the *principle of linguistic gratuity* (Wolfram 1993), and inspired by the edict ‘if knowledge is worth having, it is worth sharing’ (Cameron et al. 1992: 24). The NCLLP’s work has taken numerous shapes, ranging from publicly oriented books and audio CDs (for example, Wolfram et al. 2002), to documentary films (for example, Hutcheson 2004; Rowe and Grimes 2006), to museum exhibits and public exhibitions (for example, Vaughn and Grimes 2006), and finally to public school education materials (Reaser and Wolfram 2007).

Along with an interest in and devotion to public engagement, the NCLLP has developed a sense of duty to the data which form the backbone of its sociolinguistic research and public engagement. Thus, over the past decade, the NCLLP has increasingly been engaged in issues of corpus development, audio preservation, and, what we can broadly term, speech data management. This intersection has led to a partnership with the North Carolina State University Libraries to develop the *Sociolinguistic Archive and Analysis Project* (SLAAP; Kendall 2007, 2008).¹ SLAAP has been designed as a web-based repository and software toolkit for managing and working with sociolinguistic recordings and their related data (transcripts, variable tabulations, research notes, and so forth). It was primarily conceived of as a tool for sociolinguistic researchers, seeking to provide better management and preservation options for scholars and to develop new tools and approaches for the analysis

¹ See <https://slaap.chass.ncsu.edu>

of sociolinguistic data (Kendall 2007, 2008). In this regard, we believe that SLAAP has been, and continues to be, successful. It has an increasing user-base of researchers, houses a growing collection of data, and has aided in research projects from students' theses (for example, Kohn 2008) and dissertations (for example, Carter 2009) to published research (for example, Kendall 2013a; Mallinson and Kendall 2009; Thomas 2010; Kohn 2014) and textbooks (Thomas 2011). However, the archive and its tools have also been increasingly useful in the classroom and in other nonresearch-oriented ways and, as the project develops, its potential as a public resource becomes clearer. For example, SLAAP has enabled easier dissemination of information about dialect diversity to journalists and others, ranging from novelists and scriptwriters to dialect coaches in acting contexts.

We begin this chapter by outlining the features and design of the SLAAP website (Sect. 2). We next discuss some of the outreach projects conducted by the NCLLP (Sect. 3), and then focus in some detail on a recent book, *Talkin' Tar Heel: How our Voices Tell the Story of North Carolina* (Wolfram and Reaser 2014), which targets a public audience and makes extensive use of audio and video enhancements from our archived collection (Sect. 4). *Talkin' Tar Heel* embeds quick response codes (QRs) directly in its pages, incorporating new technologies into traditional, print formats. We end (in Sect. 5) by considering the link between our data management and outreach efforts.

2 The Sociolinguistic Archive and Analysis Project

The *Sociolinguistic Archive and Analysis Project* (SLAAP) began in earnest in 2005, originally as a digitization and preservation effort (Kendall 2007) seeking to move the NCLLP's extensive collection of audio tapes to digitized formats, to provide for future preservability and to make the recordings more accessible to researchers. Preserving and creating databases of the many recordings collected over its years of work has been an important concern for the NCLLP. The second author had amassed several large cabinets of cassette tapes over several decades of

research, both in North Carolina and from earlier in his career. These tapes were catalogued and accessible to students and colleagues through the Sociolinguistics Lab at North Carolina State University but, due to their status as analog, physical tapes in cabinets, they required local access, were hard to work with and, even worse, were degrading in quality as time passed. In its initial design, SLAAP was first envisioned as a resource specific to the NCLLP's materials (and, as indicated in Kendall 2007, it was originally titled NC SLAAP, denoting its focus on North Carolina). Over time, SLAAP has, however, grown to become a more broadly used speech data management system and recording archive. SLAAP increasingly seeks to provide a central repository for sociolinguistic recordings from outside the NCLLP and is adding large collections of non-NCLLP materials.

The specific goals behind SLAAP are multiple. On a practical level, as mentioned, the project seeks to digitize and preserve a large collection of interviews. It also aims to provide researchers with better access to and interfaces for their data through a variety of web-based features (see Kendall 2007). At a theoretical level, SLAAP questions and rethinks current linguistic and sociolinguistic conceptions of the nature of speech data, its representations, and the sorts of questions that can be asked of it (see Kendall 2008).

2.1 Design and Features of SLAAP²

SLAAP centers on a web-based archive and analytic toolset for sociolinguistic data collections, but simultaneously encompasses a broader effort to explore new approaches to storing, managing, and interacting with natural speech data. SLAAP looks to some extent like some of the other corpus development projects discussed in the recent literature (such as the ONZE Corpus discussed by Gordon et al. 2007 and the LANCHART database discussed by Gregersen 2009). However, SLAAP was developed to fill a gap in terms of sociolinguistic practice more than it sought to create a particular corpus (Kendall 2008, 2013b). In terms of Poplack's

² Parts of this section, including the specific examples (and screenshots from SLAAP), are based on Kendall (2013a).

(2007: xi) explanation of corpora design as oriented towards either *end-product* or *tool*, SLAAP was designed as a tool with no envisioned end-product. It is conceptualized as a *speech data management system* designed to house and organize an expanding collection of audio recordings. The archive continues to grow, as new recording collections are added to the database and new transcripts are developed. As of April 2015, the SLAAP digital archive contains over 4250 interviews and over 3700 hours of audio. Over 105 hours have associated time-aligned transcripts, making a transcript collection of over 1 million words. The projects housed in SLAAP are primarily from work conducted by the NCLLP in North Carolina—for example Ocracoke, NC (see Wolfram and Schilling-Estes 1995), Robeson County, NC (Wolfram and Dannenberg 1999), Hyde County, NC (Wolfram and Thomas 2002), Raleigh, NC (Dodsworth and Kohn 2012)—but also includes recordings from farther afield, for example, a study of the Northern–Midland boundary in Ohio (Thomas 2010) and research on Mexican American English in south Texas (Thomas 2015), as well as ‘historical’ recordings that form part of the sociolinguistic canon (for example, Fasold 1972; Wolfram and Christian 1976; Butters 1981). SLAAP also stores the large collection of longitudinal recordings of African American English collected by the Frank Porter Graham Project at the University of North Carolina (see Van Hofwegen and Wolfram 2010; Kohn and Farrington 2012; Kohn 2014). Most of the collections are (American) English recordings, but the archive also houses several collections of American Spanish, as well as collections from the Caribbean (for example, Reaser 2004; Myrick 2013) and even a collection of Burushaski recordings from Pakistan (Khan 2014). Finally, SLAAP increasingly includes collections from other current research groups, such as the recordings of the West Virginia Dialect Project (see, for example, Hazen 2008) and projects undertaken in Oregon (for example, McLarty et al. 2014). SLAAP is a member of OLAC, the Open Language Archives Community, and many of the collections it houses are catalogued and discoverable through the OLAC website.³

By incorporating all of these recordings into a centralized, digital repository, we have put into dialogue the many diverse collections of soci-

³ See <http://www.language-archives.org>.

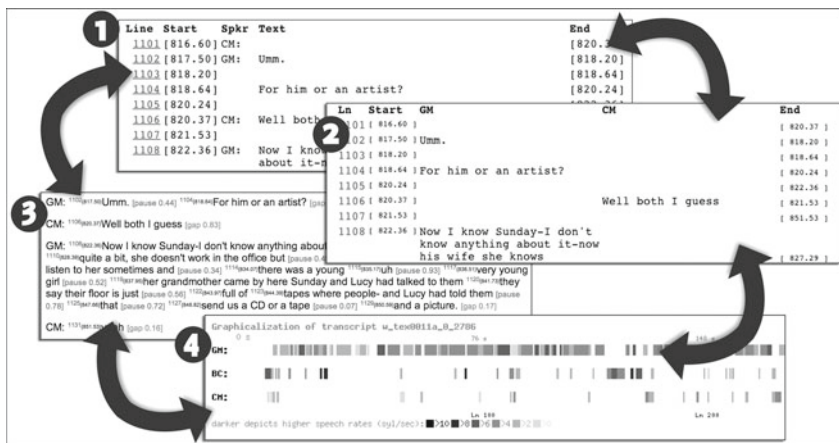


Fig. 5.1 Four presentations available in SLAAP of the same transcript data (from Kendall 2007)

olinguistic data. The descriptive metadata—the information stored about each interview, speaker, and research project—along with transcripts and researcher notes are all searchable both within and across projects. Older materials and metadata are just as easily retrieved as new materials, so tapes that formerly lay dormant for many years are now readily findable and can be put to new uses. For example, Kendall (2013a) mined transcripts and recordings from a range of projects contained in SLAAP to examine variability in speech timing features across several regional sites in the USA.

Since its beginning, we have tried to use SLAAP to explore ways that hypertext and other web technologies can enhance linguistic transcription and annotation more broadly. Our work on SLAAP has sought to make transcript information dynamic and flexible and linked to its source audio. Through SLAAP's software, the same transcript can be viewed in a *vertical format* (as in (1) in Fig. 5.1; Edwards 2001) or a *column-based format* (as in (2) in Fig. 5.1; Ochs 1979; Edwards 2001), or even in what is referred to in SLAAP as a *paragraph format* (as in (3) in Fig. 5.1). Alternatively, that same transcript can be transformed in various ways, such as into purely visual formats. The view shown in (4) of Fig. 5.1, called a *graphicalization* (Kendall 2007), displays speakers'

utterances within the complete interaction in a way that gives analysts a simple visual overview of the unfolding of the speech event. Each speaker's talk is displayed on its own tier. Shading indicates speech rate, with darker shading indicating faster speech, and pauses and speaker overlap are also accurately depicted. Analysts can generate graphicalizations with or without transcript text and can click on a passage to move to a page which provides analytic views of the transcript information (as discussed momentarily, and shown below in Fig. 5.3). Kendall (2013a) presents a second visual transformation of transcript information available in SLAAP, a so-called *Henderson graph*, which allows for the generation of quantitative metrics of a speaker's or interaction's hesitancy (see also Henderson et al. 1966; Thomas 2011: 186–7).

Transcript data in SLAAP are stored in database tables. Each transcript is a table in the database, and each line is an entry in the database table representing an utterance by a speaker. Transcripts for SLAAP are built using the TextGrid features of *Praat* (Boersma and Weenink 2015) to obtain highly accurate start- and end-times for each utterance. Each speaker is orthographically transcribed in his or her own TextGrid tier so that the temporal record accurately records the times of that specific speaker's contributions. The central unit of the transcript is the phonetic utterance—a stretch of speech bounded by pauses. Pauses are delimited separately from the speech.

Fig. 5.2 displays the *Praat* editor window for the same transcript displayed in Fig. 5.1 above. This represents the 'source' transcript before it is added to SLAAP. The example shows three utterances for the interviewee GM (the full text for the third utterance is shown by *Praat* although the actual audio, waveform, and spectrogram run off-screen to the right). The second and third tiers house the transcriptions for the two interviewers BC and CM, although in the 8-second window shown only CM speaks, with a single utterance. The interval boundaries accurately capture the start- and end-times of each utterance and, in doing so, accurately delimit the pauses. In the *Praat* window shown, the 442 millisecond pause between GM's utterance 'Umm' And 'For him or an artist?' is selected. Kendall (2013a) describes the transcription model at greater length and discusses the ways that these time-aligned, databased transcripts can be mined for sociolinguistic research purposes.

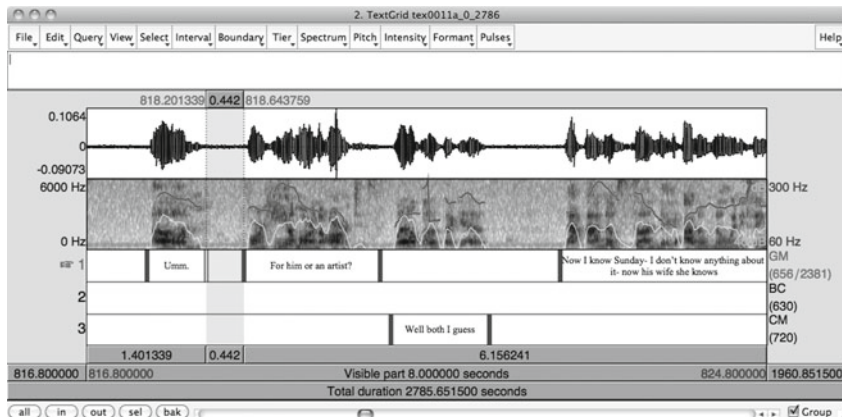


Fig. 5.2 Praat TextGrid for the transcript shown in Fig. 5.1

As this discussion illustrates, the fundamental components of SLAAP’s databased transcript model are quite simple. In such a transcript model, the only data required for a complete transcription unit are: (a) a reference as to which speaker in the interaction is speaking, (b) the utterance’s start-time, (c) an orthographic representation of the utterance, and (d) the utterance’s end-time (Kendall 2007, 2013a). Through specially designed software like SLAAP, this very simple data model is quite powerful. SLAAP creates links between the transcript data and the audio file from which the transcript is based, and phonetic software (such as *Praat* in the case of SLAAP) can be integrated into the transcript interface software to allow for real-time phonetic analysis from within the transcript. With the start- and end-times for each utterance captured in the database and a linkage maintained with the audio, much of the other information that is often tagged or coded (for example latching, overlap, pause length, and so on) is unnecessary and can be reconstructed from the audio itself.

At the same time, an approximation of standard orthography (see Chafe 1993: 34; Tagliamonte 2007: 211–15) is often sufficient for the transcript text because pronunciation features (for example, vowel qualities, *r*-vocalization, and so on) can be listened for or examined instantly via a spectrogram. This simple orthography makes the transcripts easier to read than more complex systems. The use of standard orthography also

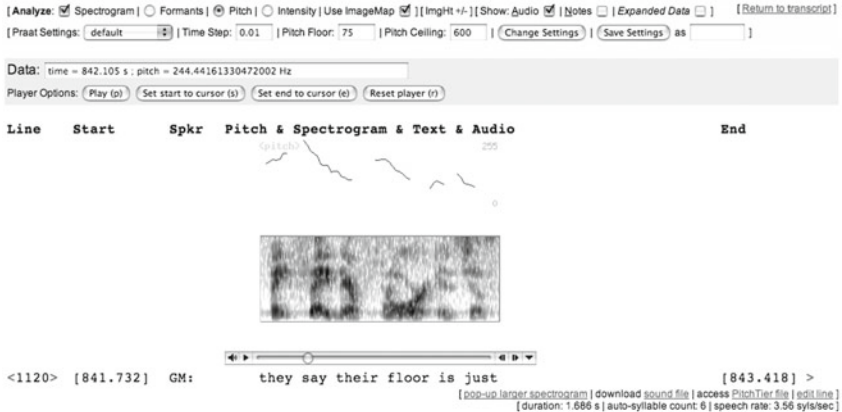


Fig. 5.3 SLAAP screenshot showing a transcript line with phonetic data

allows for easier searching and for more straightforward concordancing and other corpus-based extraction measures (Kendall 2013a). And, finally, the simple, orthographic transcription conventions and dynamic and searchable interfaces allow for flexible modes of access for a variety of different users, both linguistic analysts and nonlinguists who may be interested in the data for their content or for other purposes.

As an illustration of what can be done with this simple, databased transcript data, Fig. 5.3 shows a screenshot from the SLAAP software demonstrating an in-depth view of one transcript line. This example shows a pitch plot as well as a spectrogram, though other data views are available. Note also that the audio for the line can be listened to through an embedded audio player and that numerical data (in Fig. 5.3 this entails acoustic measurements of pitch) can be obtained at the click of the mouse. Additionally, multiple transcript lines can be displayed in this detailed format on the same page, allowing for easy comparison between utterances.

2.2 Libraries, Partnerships, and Databases

We were fortunate in the beginning phases of SLAAP to receive substantial support—moral, technical, and financial—from the Libraries at North Carolina State University (NCSU). The Libraries viewed the

preservation of the NCLLP materials as an important goal in their role as stewards of scholarly work and also viewed the project of archiving and databasing spoken language recordings as a valuable research project in library science (Kendall and French 2006; see also Smith et al. 2004). The Libraries' involvement in the project has been a crucial part of its success, and, consequently, we would like to pause briefly here to acknowledge more fully the support we have received from the NCSU Libraries. Our experiences suggest that university libraries and information technology groups are key partners in any archiving enterprise. As described in Kendall (2013b):

A persistent issue in the long-term preservation and accessibility of research recordings is the problem of institutionalization, which presents a larger hurdle than the availability of specific tools and methods or any of the technical problems of data preservation. Many sociolinguistic data collections depend on their original collector to maintain them, and many researchers create impressive websites about their work and may even maintain their own data in a web-accessible format. However, these kinds of resources take extensive time (and cost) to maintain. Traditionally, these activities have not been evaluated as a part of researchers' academic 'credit' for advancement, so we are often, in fact, disincentivized to spend the extensive and sustained effort required to ensure that our materials are accessible to others and maintained in the long term. (p. 202)

It has become increasingly clear to us that archives and projects like SLAAP cannot succeed in the long term without broader institutional support.

3 The North Carolina Language and Life Project and its Outreach Endeavors

While the preceding sections detail the preservation and analysis potential of the SLAAP system, this phase of the project might exist autonomously and be limited to its research utility. However, it is also possible to use the resources of language diversity in such a data management system with efforts to represent and exemplify speech in an outreach program

connected to public education. The following sections show how the resources of SLAAP have enhanced and enabled an outreach program focused on public education in the US state of North Carolina.

The *North Carolina Language and Life Project* (NCLLP) was established at North Carolina State University in 1993 to form an umbrella program under which research and outreach could be linked. North Carolina is a convenient state for both research and engagement since it reflects a wide variety of regional and sociocultural English dialects as well as an assortment of ancestral and immigrant languages. North Carolina's language ecology is as diverse as its physical topography and climate—the most varied of any US state east of the Mississippi River. The research-outreach mission is integral to the goals of NCLLP: (a) to gather basic research information about language varieties in order to understand the nature of language variation and change; (b) to document language varieties in North Carolina and beyond as they reflect varied cultural traditions; (c) to provide information about language differences for public and educational interests; and (d) to use research material for the improvement of educational programs about language and culture.

A state-focused model for research and engagement was adopted for political, cultural and educational reasons. Though language diversity in the USA is typically immune to state political boundaries, political governance, social and cultural institutions, as well as educational programs and standards are bounded by the state. These range from museums, the North Carolina State Fair, and other cultural organizations to the state-wide standards and objectives for the North Carolina Department of Public Instruction. Perhaps just as importantly, North Carolina is a state which takes considerable pride in its historical and cultural resources. As Wolfram and Reaser note (2014: 1), 'North Carolinians like their state a lot, and so do the 50 million yearly visitors', adding that 'North Carolina is not shy about marketing its resources as one of the top states for both living and vacationing'. In this social context, it seems advantageous to link language heritage to other cultural resources since 'language reflects where people come from, how they have developed and how they identify themselves regionally and socially' (Wolfram and Reaser 2014: 1). Although the NCLLP focuses on language variation within a political state, its goals extend further—to serve as a national model for integrating

community-based research and engagement. While a state moniker is beneficial for branding, the SLAAP archive (as discussed above), the NCLLP's research work, and the larger outreach efforts are not, of course, limited to the state of North Carolina.

3.1 Public Outreach Efforts

Since its inception, the NCLLP has been devoted to the intersection of research on language variation and change and public interests. Over the past 20 years, the NCLLP has produced 11 documentaries broadcasted on statewide, regional, and national television, constructed 6 museum exhibits, produced a half-dozen oral history audio CDs, developed a public school curriculum on language and dialect awareness, and written several trade books for popular audiences. In this section, we review some of these outreach projects, highlighting ways in which SLAAP, and our database more generally, have contributed to the efforts. Most recently, the NCLLP has published a publicly oriented book on language in North Carolina, *Talkin' Tar Heel* (Wolfram and Reaser 2014), which uses QRs to create an enhanced audiovisual experience for readers, despite its existence as a traditional print book. We focus on this most recent project in Sect. 4.

NCLLP video productions range from TV programs that have aired nationally, regionally, or on the state affiliate of the United States Public Broadcasting Service (PBS) to those produced primarily for community organizations, though a focus on local and broader-based audiences is not mutually exclusive. Examples of the former include: *Indian by Birth: The Lumbee Dialect* (Hutcheson 2001), *Mountain Talk* (Hutcheson 2004), *Voices of North Carolina* (Hutcheson 2005), *The Carolina Brogue* (Hutcheson 2008), *Spanish Voices* (Cullinan 2011), *First Language: The Race to Save Cherokee* (Cullinan and Hutcheson 2012), while the latter include: *The Ocracoke Brogue* (Blanton and Waters 1996), *Hyde Talk: The Language and Land of Hyde County* (Torbert 2002). Fortunately, high-quality video recording equipment and editing software that are portable, user-friendly, and quite affordable are now available, making video documentary production quite feasible for students and faculty. Moreover,

based on the positive responses to our programs from the public, we have even convinced the administration at NCSU to fund a full-time videographer as a core component of our research-engagement initiative.

Invariably, passages from sociolinguistic recordings archived in SLAAP are used in our documentary productions along with the collection of high-quality video recordings for each documentary. Simultaneously, the audio from video recordings captured for documentary videos have been incorporated into SLAAP. One of the challenges for the future of SLAAP is to archive and manage the hundreds of hours of video footage now collected in the production of video documentaries. At the same time, it is noteworthy that some of the video footage collected primarily for documentary production has led to new research (for example, Wolfram et al. 2014), as we have found that outreach and engagement can lead to both basic and so-called ‘engaged research’ (Reaser 2006). The relationship between research and outreach can, and should be, bilateral and synergistic rather than one-dimensional—from research to outreach.

The compilation of oral histories on CDs is yet another way that we can share the diverse voices of communities where we have conducted sociolinguistic research. Based on sociolinguistic interviews that are archived in SLAAP, and with the assistance of community members, we have compiled a number of collections of stories that reminisce, celebrate, and entertain. For example, the NCLLP staff partnered with the Ocracoke Preservation Society to produce two such compilations a decade apart, *Ocracoke Speaks* (Childs et al. 2001) and *Ocracoke Still Speaks* (Reaser et al. 2011). Staff extracted recorded stories from the archived material on SLAAP and reviewed them with community members from the Preservation Society so that the final passages for the compilation could be selected together. A similar project, *Voices of Texana*, in Texana, North Carolina, was compiled by Christine Mallinson and Becky Childs with community members in a small, isolated African American community in the Smoky Mountains (Mallinson et al. 2006), based on the archived sociolinguistic interviews in SLAAP.

The museum exhibit is another productive venue for the use of archived audio recordings from SLAAP. With community-based preservation societies and museums, it is possible to construct permanent exhibits that highlight language diversity, as well as limited-time exhibits

on history and culture, which include the voices of community residents. An exhibit titled *Freedom's Voice: Celebrating the Black Experience on the Outer Banks* (Vaughn and Grimes 2006) included images, a documentary, interactive audiovisuals, artifacts, and audio clips first recorded for sociolinguistic interviews and reappropriated oral histories to complement informational panels that highlighted African Americans' involvement in the history of coastal North Carolina. This exhibition, which ran for over a year, brought together history, culture, and language through narrating the story of the previously overlooked contributions of African Americans on coastal Carolina, particularly on Roanoke Island, the site of the 'Lost Colony'.

One of the most successful exhibits for the NCLLP is an annual booth at the North Carolina State Fair that has a yearly attendance exceeding 1 million people. Video vignettes and free dialect badges or 'buttons' as they are known in the USA (with phrases like 'bless your heart', 'dingbatter', 'I speak North Cackalacky', 'sigogglin') are always very popular with attendees. In addition, an interactive touch screen monitor allows visitors to guess the regional voices of speakers from the archival recordings. We even had attendees' pronounce different place names and added them to SLAAP for future research use. In fact, one of the most popular QRs in Wolfram and Reaser's book (2014: 11; see Sect. 4) is based on speakers recorded at the State Fair who correctly and incorrectly pronounced the names of different place names in North Carolina that symbolized insider–outsider status.

3.2 Public Education

One of the most ambitious and essential outreach programs involves the development of formal curricular materials on language diversity in the public schools. Unfortunately, formal education about dialect variation is still relatively novel and somewhat controversial, and school-based programs have still not progressed beyond a pilot stage (Reaser 2006; Sweetland 2006). The examination of dialect differences offers great potential for students to investigate the interrelation between linguistic and social diversity, including diversity grounded in geography, history, and cultural beliefs and practices.

The dialect awareness curriculum developed by Reaser and Wolfram (2007), the first program endorsed by a state Department of Public Instruction, fits in with the standard course of study for the state of North Carolina's eighth grade social studies curriculum.⁴ This language and dialect awareness program aligns with the curricular themes of 'cultures and diversity', 'historic perspectives', and 'geographical relationships' as they relate to North Carolina. In addition, the dialect awareness curriculum helps fulfill social studies competency goals such as 'Describe the roles and contributions of diverse groups, such as American Indians, African Americans, European immigrants, landed gentry, tradesmen, and small farmers to everyday life in colonial North Carolina' (Competency Goal 1.07) or 'Assess the importance of regional diversity on the development of economic, social, and political institutions in North Carolina' (Competency Goal 8.04). Students are not the only ones who profit from the study of dialect diversity. Teachers also find that some of their stereotypes about languages are challenged and that they too become more knowledgeable and informed about dialect diversity. Many of the activities are extracted from the SLAAP archive of audio recordings and video footage collected by the NCLLP. In fact, these audio and video vignettes are integral to the majority of activities as students listen to and watch clips on regional, social, and ethnic varieties integral to their workbook activities.

More information about the public engagement efforts of the NCLLP are detailed in Wolfram et al. (2008) and Wolfram (2012, 2013).

4 Digitally Reinforced Print Media: *Talkin' Tar Heel*

Writing about sociolinguistics for nonspecialized audiences is often difficult for linguists, and few seem to have creative talent for writing general books and articles for broad-based audiences. Tannen (1990, 2006) and Rickford and Rickford (2000) are notable exceptions. With varying degrees of success, the staff of NCLLP have authored several trade books aimed at these nonlinguistic audiences. For example, Wolfram and

⁴ See <http://www.ncpublicschools.org/curriculum/socialstudies/scos>.

Schilling-Estes' *Hoi Toide on the Outer Banks: The Story of the Ocracoke Brogue* (1997) is aimed at tourists and Wolfram et al.'s *Fine in the World: Lumbee Language in Time and Place* (2002) is useful for residents of the community, educators, and others curious about the language variety of the Lumbee American Indians, the largest group of American Indians East of the Mississippi River (population c.55,000) and the largest non-reservation American Indian group in the USA. However, so-called 'science accommodation'—translating highly specialized technical knowledge about science into accessible descriptions for lay people—has proven to be a rhetorical and discursive challenge for those attempting to extend their descriptions beyond the academy, including the current authors. Typically, sociolinguists have used genres readily at their disposal, such as journal articles, textbooks, invited media interviews, and occasional editorial opinion articles in newspaper and popular magazines to present their perspective.

Linguists have highly specialized analytical skills and metalanguage in the subject matter, and they need not be apologetic about this expertise. By the same token, their own attitudes associated with this knowledge can potentially lead to disrespect for and the dismissal of lay-based observations and understanding about language, and they need to be sensitive to the perspectives and ideologies of nonlinguists, including attitudes about language by highly trained colleagues in other social science and scientific fields. As Sally Johnson (2001: 592) notes, 'scientists themselves have much to learn from the reception of their ideas by those outside their area of expertise'.

Wolfram and Reaser's recent book, *Talkin' Tar Heel* (2014), represents our most recent attempt to write a book for non-linguists by focusing on language diversity within a political state. The authors note that 'after more than two decades of interviewing and recording thousands of residents and shooting hundreds of hours of video footage, we feel that we would be remiss if we did not share the rich assortment of North Carolina voices with a broader audience'. Walt Wolfram's admitted goal was 'to have my wife, Marge, read this book, listen to the audios, and see the people for herself'. Marge Wolfram, a nonlinguist, is an avid reader but finds linguistics books boring. At the same time, much like many nonlinguists, she finds language differences curious and captivating in

her everyday interactions with people. The challenge, then, is to capture and present the inherent language intrigue that language differences hold for the general public without drowning it in metalinguistic jargon.

An added feature to the print is more than 130 audiovisual enhancements. Readers get to experience the language of North Carolina as they read through the extensive use of Quick Response Codes (QRs). In Fig. 5.4,

University of North Carolina Press



FIGURE 4.4. Author Martha Pearl Villas discusses the rapid changes she has seen in Charlotte. (Photograph by Neal Hutcheson)

southern dialect: “And these No’tuhnuhs come down heuh, and we take ’em in. And befo’ you know it, it ain’t the same. It’s really not. They don’t think and ac’ like we do. Well they shu’ don’t tawk like us. They have a shahpness to theuh speech, don’t you think so? Most South-uhnuhs and all, ah mean ah feel like we have kahna a soft, melodious voice. Of cou’s, whah shouldn’t ah think it; ah don’t know any different.”²² Notice that her speech, as an elderly, upper-class resident of the city born in 1916, is characterized by *r*-lessness and *i* ungliding. And the content is permeated with the undercurrent of tension between the old and the new.

No matter how long they live in the South, transplants usually remain recognizable as outsiders to native southerners.



To view the video of this quote, visit <http://www.talkintarheel.com/chapter/4/video4-6.php>

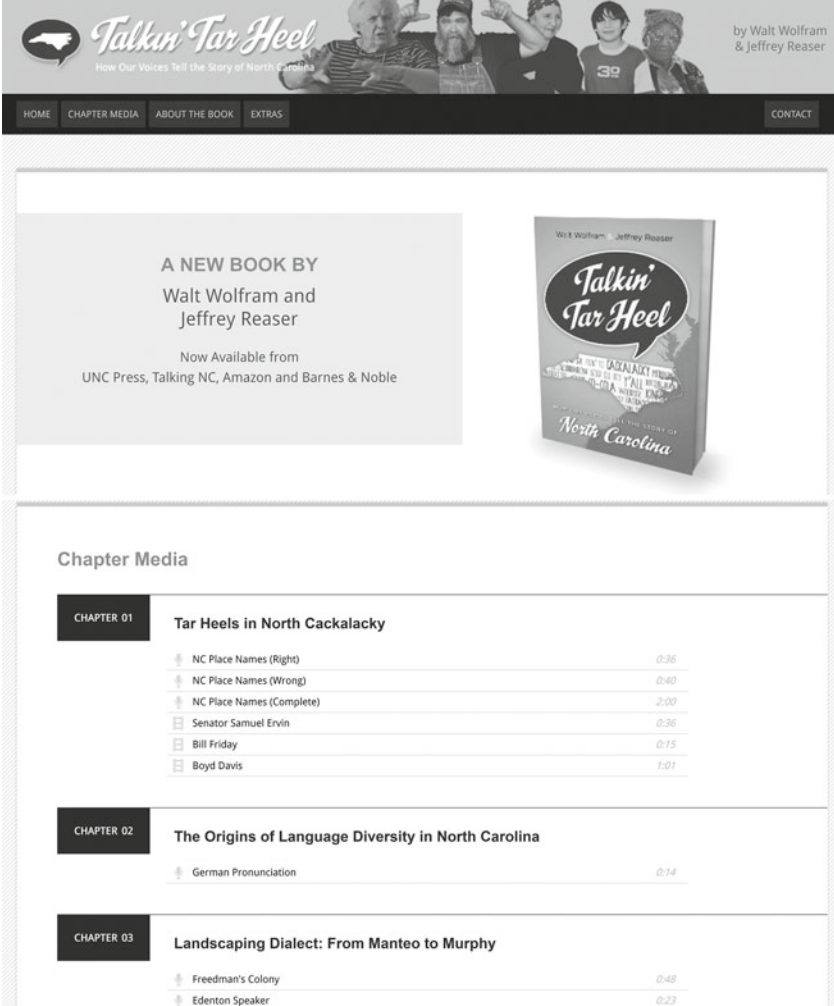


Fig. 5.4 Example of QR use with printed text

we include a copy of a page with quotes from a long-time, elderly resident of the city of Charlotte about traditional attitudes towards Northerners who move to the South. Instead of simply reading the quote in orthographically altered script, however, readers can scan the QR using any portable Internet device, like a smartphone or tablet, to go directly to the quote and see and hear the speaker for themselves. Many of the audio and video enhancements are extracted from our archive of audio and video footage to allow the voices and people to speak directly to the reader.

Though the integration of QRs is somewhat novel in the field of linguistics—*Talkin' Tar Heel* is the first book published by our publisher to use QRs and, to our knowledge the first book making extensive use of this format in linguistics—the clips seem as natural as language diversity itself. We want the general reader to experience language and dialect rather than imagine it, and accessing the enhancements is easy. Each enhancement has a brief description so that the reader knows what he or she might hear or see. All the reader has to do is navigate any web browser to the provided URL, or use a smartphone or any device with a QR reader to snap a picture of the QR code to access the media directly. In the enhanced e-Reader version of the book, these enhancements are embedded within the text itself so a reader simply has to click on the icon. The website and Chapter Media page organization for intuitive navigation of the media clips on the website are displayed in Fig. 5.5.

A number of outreach opportunities have derived from the publication of the book. For example, in the months following the book's publication, Wolfram and Reaser gave more than 20 book readings at independent and chain bookstores around the state where they invariably demonstrated the integration of the audiovisual QR enhancements in the book. They have also been invited to speak at numerous civic organizations, preservation societies, and other state- and community-based events, and conducted television and radio interviews where they demonstrated how the audiovisual examples are integrated into the print text. These opportunities have extended outreach in ways that even they did not imagine. One customer was so enthusiastic about the potential of the use of QRs for public school education that he purchased \$2500 in books for teachers and contributed \$5000 to develop education materials related to the book for middle-school students. Responses such as these have exceeded our expectations



The screenshot shows the website for the book "Talkin' Tar Heel: How Our Voices Tell the Story of North Carolina" by Walt Wolfram and Jeffrey Reaser. The website features a navigation bar with links for HOME, CHAPTER MEDIA, ABOUT THE BOOK, EXTRAS, and CONTACT. The main content area includes a promotional banner for the book, a 3D rendering of the book cover, and a "Chapter Media" section. The "Chapter Media" section is organized into three chapters, each with a list of audio segments and their durations.

Chapter Media

CHAPTER 01 Tar Heels in North Cackalacky

NC Place Names (Right)	0:36
NC Place Names (Wrong)	0:40
NC Place Names (Complete)	2:00
Senator Samuel Ervin	0:36
Bill Friday	0:15
Boyd Davis	1:01

CHAPTER 02 The Origins of Language Diversity in North Carolina

German Pronunciation	0:14
----------------------	------

CHAPTER 03 Landscaping Dialect: From Manteo to Murphy

Freedman's Colony	0:48
Edenton Speaker	0:23

Fig. 5.5 Website and Chapter Media for *Talkin' Tar Heel*

in terms of the potential for research and outreach to enable each other. Furthermore, they underscore the potential for interdisciplinarity that includes not only the collaboration of academic fields but technical and professional collaboration as well. In our outreach efforts, we work with designers, producers, programmers, marketing experts, and other professionals who enable both research and engagement.

5 Conclusion: Engagement and Data Management

On the face of it, the connection between the NCLLP's engagement work on the one hand and the archival and data management work of SLAAP may seem like a contrived, tenuous one. It is certainly true that the primary connection is through the data, as both endeavors pivot on the core data. The data that form the backbone and *raison d'être* for SLAAP are necessary for any empirically based research effort, but they are also critical for dialect education and outreach efforts. In fact, the two sets of activities focused on in this chapter are just two points of a troika, with the third being the center of interest for most academic linguists—the study of language (and language variation and change) itself. While data management was not discussed in Wolfram's (1993) original formulation of the principle of linguistic gratuity, which stated that 'investigators who have obtained linguistic data from members of a speech community should actively pursue positive ways in which they can return linguistic favors to the community', or much of the other early literature on research ethics (Labov 1982; Cameron et al. 1992) it is nonetheless the case that being good stewards of the data we collect is a way to 'return linguistic favors' to the communities we study. It is indeed possible to enrich both the communities from whom we obtain our data and to preserve that data for the study of and celebration of language variation.

References

Books, Articles, CDs and DVDs

- Blanton, Phyllis and Karen Waters, producers. 1996. *The Ocracoke Brogue*. Raleigh, NC: North Carolina Language and Life Project.
- Butters, Ronald R. 1981. Unstressed vowels in Appalachian English. *American Speech* 56(2): 104–110.
- Cameron, Deborah, Elizabeth Frazer, Harvey Penelope, M.B.H. Rampton, and Kay Richardson. 1992. *Researching Language: Issues of Power and Method*. London/New York: Routledge.

- Carter, Phillip. 2009. *Speaking Subjects: Language, Subject Formation, and the Crisis of Identity*. Doctoral Dissertation. Duke University.
- Chafe, Wallace. 1993. Prosodic and functional units of Language. In *Talking Data: Transcription and Coding in Discourse Research*, eds. Jane Edwards, and Martin D. Lampert, 33–43. Hillsdale, NJ: Lawrence Erlbaum.
- Childs, Becky, Walt Wolfram and Ellen Fulcher Cloud, producers. 2001. *Ocracoke Speaks*. Raleigh, NC: North Carolina Language and Life Project.
- Cullinan, Danica, producer. 2011. *Spanish Voices*. Raleigh, NC: North Carolina Language and Life Project.
- Cullinan, Danica and Neal Hutcheson, producers. 2014. *First Language: The Race to Save Cherokee*. Raleigh, NC: North Carolina Language and Life Project.
- Dodsworth, Robin, and Mary Kohn. 2012. Urban rejection of the vernacular: The SVS undone. *Language Variation and Change* 24(2): 221–245.
- Edwards, Jane. 2001. The transcription of discourse. In *Handbook of Discourse Analysis*, eds. Deborah Tannen, Deborah Schiffrin, and Heidi Hamilton, 321–348. Malden, MA/Oxford: Blackwell.
- Fasold, Ralph W. 1972. *Tense Marking in Black English: A Linguistic and Social Analysis*. Washington, DC: Center for Applied Linguistics.
- Gordon, Elizabeth, Margaret Maclagan and Jennifer Hay. 2007. The ONZE Corpus. In *Creating and Digitizing Language Corpora. Volume 2: Diachronic Databases*, Joan C. Beal, Karen P. Corrigan and Hermann L. Moisl (eds), 82–104. New York/Basingstoke: Palgrave Macmillan.
- Gregersen, Frans. 2009. The data and design of the LANCHART study. *Acta Linguistica Hafniensia* 41: 3–29.
- Hazen, Kirk. 2008. (ING): A vernacular baseline for English in Appalachia. *American Speech* 83(2): 116–140.
- Henderson, Alan, Frieda Goldman-Eisler, and Andrew Skarbek. 1966. Sequential temporal patterns in spontaneous speech. *Language and Speech* 9: 207–216.
- Hutcheson, Neal, producer. 2001. *Indian by Birth: The Lumbee Dialect*. Raleigh, NC: North Carolina Language and Life Project.
- Hutcheson, Neal, producer. 2004. *Mountain Talk*. Raleigh, NC: North Carolina Language and Life Project.
- . 2005. *Voices of North Carolina*. Raleigh, NC: North Carolina Language and Life Project.
- Hutcheson, Neal. 2008. *The Carolina Brogue*. Raleigh, NC: North Carolina Language and Life Project.

- Johnson, Sally. 2001. Who's misunderstanding whom? Sociolinguistics debate and the media. *Journal of Sociolinguistics* 5: 591–610.
- Kendall, Tyler. 2007. The Sociolinguistic Archive and Analysis Project: empowering the sociolinguistic archive. *Penn Working Papers in Linguistics* 13(2): 15–26.
- . 2008. On the history and future of sociolinguistic data. *Language and Linguistics Compass* 2(2): 332–351.
- . 2013a. *Speech Rate, Pause, and Sociolinguistic Variation: Studies in Corpus Sociophonetics*. New York/Basingstoke: Palgrave Macmillan.
- . 2013b. Data preservation and access. In *Data Collection in Sociolinguistics: Methods and Applications*, Christine Mallinson, Becky Childs and Gerard Van Herk (eds), 195–205. New York: Routledge.
- Kendall, Tyler and Amanda French. 2006. Digital audio archives, computer-enhanced transcripts, and new methods in sociolinguistic analysis. Paper presented at Digital Humanities (ALLC/ACH) 2006. Paris, France. July.
- Khan, Samina. 2014. *Burushaski Recordings*. Raleigh, NC: North Carolina Language and Life Project.
- Kohn, Mary. 2008. *Latino English in North Carolina: A Comparison of Emerging Communities*. Masters Thesis. North Carolina State University.
- . 2014. “*The Way I Communicate Changes but How I Speak Don't*”: *A Longitudinal Perspective on Adolescent Language Variation and Change*. Publication of the American Dialect Society 99. Durham: Duke University Press.
- Kohn, Mary, and Charlie Farrington. 2012. Speaker normalization: evidence from longitudinal child data. *Journal of the Acoustical Society of America* 131(3): 2237–2248.
- Labov, William. 1982. Objectivity and commitment in linguistic science. *Language in Society* 11: 165–201.
- Mallinson, Christine, Becky Childs, and Zula Cox. 2006. *Voices of Texana*. Texana, NC: Texana Committee on Community History and Preservation.
- Mallinson, Christine, and Tyler Kendall. 2009. “The way I can speak for myself”: the social and linguistic context of counseling interviews with African American adolescent girls in Washington, DC. In *African American Women's Language*, ed. S.L. Lanehart, 110–126. Newcastle upon Tyne: Cambridge Scholars Press.
- McLarty, Jason, Charlie Farrington, and Tyler Kendall. 2014. Perhaps we used to but we don't anymore: the habitual past in Oregonian English. *Penn Working Papers in Linguistics* 20(2): 111–119.

- Myrick, Caroline. 2013. Big-time variation in small-time communities: sociolinguistic diversity on Saba. Paper presented at the Southeastern Conference on Linguistics (SECOL) 80. Spartanburg, SC.
- NC [North Carolina] Standard Course of Study. Social Studies 2006: Eighth grade North Carolina: creation and development of the state. Online document. <http://www.ncpublicschools.org/curriculum/socialstudies/scos/2003-04/050eighthgrade> (accessed 22 May 2014).
- Ochs, Elinor. 1979. Transcription as theory. In *Developmental Pragmatics*, eds. Elinor Ochs, and Bambi Schieffelin, 43–72. New York: Academic Press.
- Poplack, Shana. 2007. Foreward. In *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*, Joan C. Beal, Karen P. Corrigan and Hermann L. Moisl (eds), ix–xiii. New York/Basingstoke: Palgrave Macmillan.
- Reaser, Jeffrey. 2004. A quantitative sociolinguistic analysis of Bahamian copula absence: morphosyntactic evidence from Abaco Island, the Bahamas. *Journal of Pidgin and Creole Languages* 19: 1–40.
- . 2006. *The Effect of Dialect Awareness on Adolescent Knowledge and Attitudes*. Durham: Duke University dissertation.
- Reaser, Jeffrey, and Walt Wolfram. 2007. *Voices of North Carolina: Language and Life from the Atlantic to the Appalachians*. Raleigh, NC: North Carolina Language and Life Project.
- Reaser, Jeffrey, Paula Dickerson Bunn, Walt Wolfram, DeAnna Locke, Chester Lynn, and Phillip Howard. 2011. *Ocracoke Still Speaks: Reflections Past and Present*. Raleigh, NC: North Carolina Language and Life Project.
- Rickford, John R., and Russell J. Rickford. 2000. *Spoken Soul: The Story of Black English*. New York: John Wiley & Sons.
- Rowe, Ryan and Drew Grimes, producers. 2006. *This Side of the River*. Raleigh, NC: North Carolina Language and Life Project.
- Smith, Abby, David Allen, and Karen Allen. 2004. *Survey of the State of Audio Collections in Academic Libraries*. Washington, DC: Council on Library and Information Resources.
- Sweetland, Julie. 2006. *Teaching Writing in the African American Classroom: A Sociolinguistic Approach*. Palo Alto, CA: Stanford University dissertation.
- Tagliamonte, Sali. 2007. Representing real language: consistency, trade-offs and thinking ahead! In *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*, Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl (eds), 205–240. New York/Basingstoke: Palgrave Macmillan.
- Tannen, Deborah. 1990. *You Just Don't Understand: Women and Men in Conversation*. New York: Ballantine.

- . 2006. *You're Wearing That? Understanding Mothers and Daughters in Conversation*. New York: Random House.
- Thomas, Erik R. 2010. A longitudinal analysis of the durability of the Northern-Midland dialect boundary in Ohio. *American Speech* 85(4): 375–430.
- . 2011. *Sociophonetics: An Introduction*. New York/Basingstoke: Palgrave Macmillan.
- Thomas, Erik R. 2015. What a swarm of variables tells us about the formation of Mexican American English. Paper presented at LAVIS IV. Raleigh, North Carolina State University.
- Torbert, Benjamin, producer. 2002. *Hyde Talk: The Language and Land of Hyde County, NC*. Raleigh, NC: North Carolina Language and Life Project.
- Van Hofwegen, Janneke, and Walt Wolfram. 2010. Coming of age in African American English: a longitudinal study. *Journal of Sociolinguistics* 14(4): 427–455.
- Vaughn, Charlotte, and Drew Grimes. 2006. *Freedom's Voice: Celebrating the Black Experience on the Outer Banks*. Manteo, NC: Outer Banks History Center and North Carolina Language and Life Project.
- Wolfram, Walt. 1993. Ethical considerations in language awareness programs. *Issues in Applied Linguistics* 4: 225–255.
- . 2012. In the profession: connecting with the public. *Journal of English Linguistics* 40(1): 111–117.
- . 2013. Community commitment and responsibility. In *The Handbook of Language Variation and Change*, 2 edn, eds. J.K. Chambers, and N. Schilling, 557–576. Malden: Wiley-Blackwell.
- Wolfram, Walt, and Donna Christian. 1976. *Appalachian Speech*. Washington, DC: Center for Applied Linguistics.
- Wolfram, Walt, and Clare Dannenberg. 1999. Dialect identity in a tri-ethnic context: the case of Lumbee American Indian English. *English World Wide* 20: 179–216.
- Wolfram, Walt, Clare Dannenberg, Stanley Knick, and Linda Oxendine. 2002. *Fine in the World: Lumbee Language in Time and Place*. Raleigh, NC: North Carolina State University.
- Wolfram, Walt, Jaclyn Daugherty, and Danica Cullinan. 2014. On the (in) significance of English language variation: Cherokee and Lumbee English in comparative perspective. *Penn Working Papers in Linguistics* 20(2): 199–208.
- Wolfram, Walt, and Jeffrey Reaser. 2014. *Talkin' Tar Heel: How Our Voices Tell the Story of North Carolina*. Chapel Hill, NC: University of North Carolina Press.

- Wolfram, Walt, Jeffrey Reaser, and Charlotte Vaughn. 2008. Operationalizing linguistic gratuity: from principle to practice. *Language and Linguistics Compass* 2(6): 1109–1134.
- Wolfram, Walt, and Natalie Schilling-Estes. 1995. Moribund dialects and the endangerment canon: the case of the Ocracoke brogue. *Language* 71: 696–721.
- . 1997. *Hoi Toide on the Sound Side: The Story of the Ocracoke Brogue*. Chapel Hill, NC: University of North Carolina Press.
- Wolfram, Walt and Erik R. Thomas. 2002. *The Development of African American English*. Oxford, UK/Malden, MA: Blackwell.

Websites, Software and Online Resources

- Boersma, Paul and David Weenink. 2015. *Praat: doing phonetics by computer*. Version 5.4.15. <http://www.praat.org> (accessed 9 August 2015).
- North Carolina Language and Life Project (NCLLP). <http://www.ncsu.edu/linguistics/ncllp> (accessed 17 April 2015).
- Open Language Archives Community (OLAC) <http://www.language-archives.org> (accessed 17 April 2015).
- Sociolinguistic Archive and Analysis Project (SLAAP). <https://slaap.chass.ncsu.edu> (accessed 17 April 2015).
- Talkin' Tar Heel: How Our Voices Tell the Story of North Carolina. <http://www.talkintarheel.com> (accessed 17 April 2015).

6

Roswell Voices: Community Language in a Living Laboratory

William A. Kretzschmar, Jr

1 Introduction

The idea of partnerships between academic linguists and the communities in which they work has become much more complex over the years. Walt Wolfram's idea of linguistic gratuity, that 'investigators who have obtained linguistic data from members of a speech community should actively pursue ways in which they can return linguistic favors to the community' (Wolfram 1993: 227) challenged sociolinguists to make their work a two-way process. The idea of public corpora (for example, Kretzschmar et al. 2006a) has been promoted as one way to give something back to the community and at the same time allow other linguists besides original investigators to access the data. Wolfram and students such as

W.A. Kretzschmar, Jr (✉)
University of Georgia, Athens, GA 30602, United States
University of Glasgow, Glasgow G12 8QQ, United Kingdom
University of Oulu, Pentti Kaiteran katu 1, 90014 Oulu, Finland

Jeff Reaser have developed programs for schools (for example, Wolfram et al. 2008). Now, the idea of partnerships has expanded beyond giving something back to the community or making data available to the public: communities and linguists can work together in living laboratories, in which both sides take an active role in both community development and linguistic research. Communities as living laboratories can address economic and cultural issues as well as those in research and education. This chapter describes activities in Roswell, Georgia, that exploit such a partnership, including membership in the European Union's Living Laboratory Network.

2 Beginnings

The *Roswell Voices* project began in 2002 as a partnership between a team of researchers at the University of Georgia (UGA) and the Roswell Folk and Heritage Bureau, a division of the Roswell Convention and Visitors Bureau (CVB), to collect conversational interviews to document language and life in the community. Growth accelerated as the community became a suburb of Atlanta, and now Roswell has blossomed as an 'edge city', where residents typically live and work in Roswell, and believe themselves to be residents of the city itself as opposed to Atlanta. Two representatives from the CVB had heard of Walt Wolfram's work with North Carolina towns and wondered whether there might be a distinctive way of talking in Roswell as Wolfram had described community speech in places like Ocracoke. Wolfram had referred them to me as somebody more local. I agreed to begin an investigation of local speech in Roswell, not because I thought that there was likely to be an accent like those that Wolfram had found in isolated communities, but because Roswell is perfectly positioned for a study of the effects on language of the demographic change that has characterized America at the end of the twentieth century: first suburbanization and later the emergence of former suburbs as their own independent communities. Two colleagues then at UGA, Sonja Lanehart and Bridget Anderson, joined me to plan and conduct interviews.

The site of Roswell was in Cherokee territory, but little remains there now of these Native Americans after the 'Trail of Tears' in 1838. In 1839 the

Roswell settlement was founded with a textile mill on the Chattahoochee River, and developed as a mill town with three distinct social groups: the prominent families who lived in nice houses, the mill workers who lived in cottages and apartments, and slaves (see Heath's *Ways with Words* (1983), and counterpoint from another Georgia town in McNair (2005), which describes language realignment between mill workers and farmers). The town achieved a population of 1000 only between 1870 and 1880, and did not get to 2000 until about 1950. However, at the end of the mill period in 1975, the population had climbed to 15,000. It was 23,000 in 1980, more than doubling to 48,000 in 1990, and nearly doubling again to approximately 100,000 at the present time. The area of Roswell has grown from a village on the Chattahoochee to encompass a land area of about 39 square miles. As of the 2000 census, about half of Roswell residents had a household income over \$100,000, and the median price for a house there was about \$200,000. Roswell was over 80 per cent white and only 8.5 per cent black, with a growing Hispanic population that had already reached 10.5 per cent by the year 2000. The process of demographic change carried Roswell beyond the status of 'suburb' to become more like a city in its own right, an 'Edge City' (Zelinsky 1992: 166, who cites Garreau 1991). Many people still commute from edge cities to work in the central city, but it has come to have an exurban identity separate from that of the central city, and its own business core. Roswell's largest employer is Kimberly-Clark, a paper goods company with annual revenue over \$15 billion, and there are over 5000 businesses now registered in the city. The Roswell CVB is a tireless promoter of business and cultural events in the community, promoting the town with its several historical, large churches as a center for weddings, and with events such as the large annual Memorial Day celebration and an autumn Arts Festival.¹ Unlike the isolated towns in the region sometimes studied by sociolinguists (as for example by Wolfram's team in North Carolina), the history of Roswell and its lively modern character exemplify the main line of demographic change in America, and thus offers an opportunity to study language change in that context.

¹ See <http://visitroswellga.com>.

From the beginning, the Roswell–UGA partnership respected the interests of both sides. The CVB was interested in interviews with the oldest surviving generation in Roswell, some of them iconic figures in the community, and we interviewed those people. We created a guided conversational interview protocol that allowed for elicitation of a great deal of oral history from participants, all of whom were willing volunteers. After collecting detailed demographic information, we asked about churches and related activities like weddings, about daily life in the community including chores and foodways, about schools and integration, and about urbanization in Roswell and its relation to Atlanta. We then asked about historic buildings and institutions in Roswell, and about any key moments in the development of Roswell that the participants wanted to talk about. We learned from the oldest living generation about the first cars in Roswell, about when electricity came to the town, and about how to smoke meat and how to plant vegetables ‘by the signs’ in traditional Southern subsistence farming culture. At the end of the conversational part of the interview, we used fixed-format elicitation to document pronunciation by having participants read a set of cue cards with different phonetic segments in different environments, and to document the lexicon by asking a set of 24 questions about what participants called things, like ‘the store that is also located where you get gas that sells items like milk and beer and small food items and batteries’. Of course the fixed-format elicitation served linguistic ends, so that we got comparable data from each speaker along with the conversational data about their speech. Besides the oldest living generation, we also interviewed their children and their children’s children—three generations. The CVB did not start out with an interest in the younger speakers but always supported the UGA team’s wishes to find and interview them.

3 Phases I and II

Working with only minimal funding for a part-time graduate assistant and volunteers, the first two phases of *Roswell Voices* interviewed the oldest living residents in both the white and the historic African American communities in Roswell along with speakers from the middle generation

and young adults. In the first phase (2002–4) the UGA team conducted 22 interviews with Roswell residents, 9 with white speakers and 9 with African American speakers of the oldest and middle generations, plus 4 interviews with white young adult speakers. We first published our findings in a pamphlet called *Roswell Voices* (Kretzschmar et al. 2004), which featured a short introduction about the many voices to be found in Roswell (not just the one distinctive accent the CVB had originally hoped to find), followed by short vignettes selected from the many interviews up to that time. The pamphlet, which was available for sale in the CVB office, was accompanied by a CD so that users could listen to the voices in Roswell as well as read them in print, and the CD could be heard in the CVB office next to large display boards about the *Roswell Voices* project. For this work the UGA Team was awarded the President's Award by the Roswell CVB in 2004, in association with the sesquicentennial celebration of the community. Thereafter, the UGA team, now consisting just of Kretzschmar and part-time graduate assistants, continued to work with minimal funding to collect interviews. In the second phase (2005–6), 29 additional interviews were conducted with Roswell residents of both races and all ages. We published a second pamphlet, *Roswell Voices, Phase 2* (Kretzschmar et al. 2006b), which again featured vignettes of speakers and a CD so that users could listen as well as read. While the second pamphlet did include stories from the oldest generation, it included more stories from younger speakers about topics like race relations and making one's way in the community. These interviews expanded the scope of the Roswell project beyond the mainly older population of Phase I so that we had a more balanced picture of the modern community. Again, the pamphlet was available for sale in the CVB office, and was accompanied by a large display. These pamphlets and the display boards in the CVB office constituted a tangible result of the partnership for the community. Corporate partners of the CVB had provided the costs for printing and duplication of CDs, so proceeds supported the CVB. Participants all received a copy on CD of their complete interviews, and these have been popular, especially in the families of the oldest generation, some of whom, as one might expect, have now passed away. Subsequently, a number of audio files from the pamphlets have been mounted by the CVB on its website for the Folk

Roswell Folk & Heritage Bureau

HOME KEEPERS LANGUAGE & LIFE ROSWELL MILLS HISTORY STORYTELLING THE CHEROKEE CONTACT US

Language & Life Division

The Roswell Folk & Heritage Bureau is delighted to provide a speaker for your group, in order to encourage awareness and preservation of our traditional culture, customs, and folkways. To arrange a program, please call (770) 640-3253.

The Roswell Folk & Heritage Bureau works closely with experts in Georgia and throughout the South to pull together a dynamic approach to this fascinating subject. It is our belief that one can't appropriately understand an area without understanding the cultures that have blended to form its dialects, traditions, and customs.

Roswell Voices

A Project of Roswell Folk & Heritage Bureau Language & Life Division

The dialect with which we speak is a key to knowing our ancestors. Each area of the country has its own regional dialect. These dialects, in their truest forms are rapidly changing.

The Roswell Folk & Heritage Bureau, Language & Life Division, documents Roswell's community language through Roswell Voices, an oral history and dialect awareness program. Working with Bill Kretzschmar (PhD), professor of English and Linguistics at the University of Georgia, we have documented many long-time Roswell residents, their dialects and their stories. We are also documenting community change by comparing these interviews with interviews from younger generations. The Roswell Folk & Heritage Bureau, in cooperation with the team from the University of Georgia, has produced two booklets with accompanying compact disk of the Roswell Voices Project.

Roswell Voices – English

Aubrey Gentry

Barbara Horton

Roswell Voices – Hispanic

The American Dream (El sueño americano)

Es una vida la puedes vivir bien

Fig. 6.1 Roswell Voices on the Roswell Folk and Heritage Bureau website

and Heritage Bureau (Fig. 6.1).² Information on the page says that ‘[t]he dialect with which we speak is a key to knowing our ancestors. Each area of the country has its own regional dialect. These dialects, in their truest forms are rapidly changing’, and further that ‘we have documented many long-time Roswell residents, their dialects and their stories. We are also documenting community change by comparing these interviews with interviews from younger generations’. Thus the CVB has retained its historical interests, but now embraces the interest of the UGA team in demographic change and its impact on speech.

² See <http://roswellheritage.com/language-life>.

4 Spanish in Roswell

A third stage of interviews occurred somewhat later, between 2011 and 2013. A talented undergraduate at UGA, Anna Wilson, decided to learn about sociolinguistics by working in Roswell with its new Spanish-speaking residents, who by that time amounted to nearly a quarter of the community. This was not the first time that UGA undergraduates had taken an interest in Roswell: I had introduced many of them to the community, even taking groups there to see the town and its significant landmarks so that they could get a feel for the vignettes we studied in small seminars back on the UGA campus. The CVB facilitated these visits, for example by arranging tours of Bulloch Hall, a historic (1839) home from the founding of Roswell that later came to have connections with the Roosevelt family. Sociolinguistics students at UGA got practical fieldwork experience by helping with Roswell interviews, part of the ‘service learning’ model currently growing in American universities whereby university courses can help to accomplish community goals. Josh Dunn conducted nine more interviews with speakers of the youngest generation after Phase 2, as part of an Honors thesis on developing trends in Roswell speech. Wilson, too, completed an Honors thesis on Roswell speech, after having received credit and funding from UGA’s undergraduate research program (CURO). As for the newer Spanish population, it took time for Wilson to find her way into the Roswell Latino community, but with the help of the CVB she was able to make the necessary contacts and conducted 26 interviews in Spanish. The CVB again solicited support from local businesses to print *Roswell Voices, Phase 3* (Wilson 2013), which again features short vignettes in text and in audio on CD. As the front matter says: ‘You can hear the Spanish voice of Roswell on the accompanying CD, while you can also read a transcript in Spanish and a translation in English’. The CVB put seven of these vignettes, the first in English about ‘The American Dream’ and the rest in Spanish, on the Folk and Heritage Bureau website.³ Again, the CVB supported the interests of the UGA team, and embraced them as part of the history and culture of the

³ See <http://roswellheritage.com/language-life>.

community. Indeed, Wilson's work in the Latino community prompted the CVB to explore ways to make better connections with this part of its economic base. The partnership has served well to build a picture of Roswell culture, both historically and in the present day, that the CVB can use as part of its mission to promote the community and that UGA has used for educational goals.

5 Academic Research on Roswell Speech

The academic research side of the partnership has not been neglected. A key resource at UGA is the new Atlas website built in association with the UGA Main Library,⁴ to which the Roswell materials have been appended and archived as a separate unit. Interview materials have been made secure and initial full-text transcriptions have been made of 40 interviews, and 57 interviews have been made available online.⁵ Table 6.1 provides an idea of the social situation for many of the Phase I and II speakers. There has not yet been an opportunity for in-depth acoustical phonetic processing to exploit their considerable value for linguists, but preliminary analysis (Anderson 2005; Kretzschmar 2005; Kretzschmar et al. 2007) has indicated that there are strong generational shifts in pronunciation, such as the failure of the younger generation to repeat their grandparents' and parents' loss of -r in unstressed syllables, diphthongal pronunciation of *dog*, [æu] in words like *house*, or occasional loss of postvocalic -r in stressed syllables; that the oldest African Americans in Roswell had an upland Southern speech type (parallel to findings in North Carolina by Wolfram's group: Wolfram and Thomas 2002; Childs and Mallinson 2004; Hilliard and Carpenter 2004; Rowe and Kendall 2004); that features of 'national' African American speech were heard from the middle and younger generations; and that cultural change in the community generally frustrated attempts to elicit lexical evidence that might be compared to earlier Linguistic Atlas interviews from the same and neighboring counties (from LAGS, Pederson 1986–92). Acoustic phonetic analysis

⁴ See <http://www.lap.uga.edu>.

⁵ See <http://www.lap.uga.edu/Projects/ROSWELL/Speakers>.

Table 6.1 Phase I and Phase II interviews

Gender	Year of birth	Education level	Religion	Primary occupation	Race	Date of interview	Interviewer
F	na	na	na	na	AA	na	Lanehart and Childs
F	na	na	na	na	AA	na	Lanehart and Childs
F	1923	na	na	na	W	2/17/2006	Andres
M	na	na	na	na	W	11/6/2003	Votta
F	1915	BA in edu, part of masters	B		W	2/17/2006	Andres
M	1984	in college in 2006	B	na	AA	4/19/2006	Johnson
na	na	na	na	na	na	na	Anderson and Childs
M	1923	10th grade	na	Cabinet maker	W	2/13/2006	Andres
na	na	na	na	na	W	na	Anderson and Childs
M	1950	na	UM	Exterminator	W	1/20/2006	Kretzschmar
M	1922	na	na	na	W	2/20/2006	Lanehart and Childs
M	na	na	na	na	W	2/20/2006	Andres
M	1928	11th grade	B	Woodworking business	W	2/6/2006	Andres
na	na	na	na	na	W	na	Anderson & Childs
F	1939	1 Year of college	B	na	W	3/3/2006	Andres
na	na	na	na	na	W	na	Anderson and Childs
F	1935	College	B	Painting contractor	W	1/20/2006	Kretzschmar
M	na	na	na	na	na	na	na
M	1919	College	B	Baptist pastor	W	2/13/2006	Andres
M F	na	na	na	na	W	1/30/2006	Andres
M	na	na	na	na	na	11/11/2003	Kretzschmar
M	1936	2 Years of college	na	Automotive work	W	2/3/2006	Andres
M	na	na	na	na	W	na	Anderson and Childs
F	1943	9th grade	na	Breakfast manager in a hotel	W	3/17/2006	Andres
M	1964	High school	na	General manager in a hotel	W	3/18/2006	Andres

(continued)

Table 6.1 (continued)

Gender	Year of birth	Education level	Religion	Primary occupation	Race	Date of interview	Interviewer
M	1967	College	B	Minister and truck driver	AA	4/19/2006	Johnson
M	1915	7th grade	na	Cooking, domestic	AA	10/15/2003	Lanehart and Childs
F	1920	Read & write	na	na	AA	10/22/2003	Lanehart and Childs
F	1947	College	B	Office manager	AA	4/1/2006	Johnson
F	1935	High school	B	Secretary	W	2/3/2006	Votta
F	1917	9th grade	na	Domestic, Georgia Power	AA	10/22/2003	Lanehart and Childs
M F	na	na	na	na	AA	11/21/2003	Lanehart
F	1910	College	na	na	na	na	Anderson and Childs
F	1925	na	na	na	W	2/17/2006	Andres
M F	na	na	na	na	na	na	Votta
F	na	na	E	na	na	2/3/2006	Andres
F	1957	High school	B	Diet clerk at hospital	W	3/31/2006	Andres
F	1980	High school	CR	Dance instructor	AA	4/8/2006	Johnson
M	1920	High school	B	Grocer, school bus driver, fire chief	AA	4/8/2006	Johnson
F	1983	College	J	Marketing specialist	W	1/30/2006	Votta
na	na	na	na	na	W	2/3/2006	Votta
na	na	na	na	na	na	na	Anderson and Childs
F	1953	College	B	Consultant	na	na	na
M	na	na	na	na	AA	4/1/2006	Johnson
M	1972	In college in 2006	CR	Customer service	AA	na	Lanehart
M	1973	College	B	Pastor	AA	4/8/2006	Johnson
M	na	na	na	na	AA	11/21/2003	Lanehart
F	1983	In college, has BA	J	Student	na	na	Anderson and Childs
M	1930	College		Airforce	W	10/15 and 10/27/2003	Kretzschmar
F	1936	2½ Years in college	B	Teacher's assistant	W	1/30/2006	Votta
F	1984	College	C	Works at visual arts center	AA	4/1/2006	Johnson
					W	11/6/2003	Kretzschmar

from four pairs of speakers was included in the 2007 LSA Symposium on Vowel Phonology and Ethnicity (Andres and Votta 2007), and analysis from six pairs of speakers was included in the volume published from that event (Andres and Votta 2009). Results of this study show high variability between speakers, and do not support either the Southern Shift hypothesis or the national African American system proposed by Bailey and Thomas (1998): no systematic reversal of *i/I* or *eI/ε*, and no systematic glide shortening. The findings were also different from Kretzschmar's (2015) in Atlanta, where African Americans do reverse *eI/ε*. However, Andres and Votta did find small groups who shared isolated shift characteristics (three men with *cotI/caught* merger, two women and one man with *eI/ε* values approaching the shift). Dunn (2008) presented impressionistic evidence from our youngest cohort of Roswell English speakers that shows high variability in their use of traditional features of Southern speech. This finding was amplified and contextualized in three papers (Kretzschmar and Dunn 2010; Kretzschmar et al. 2011; Kretzschmar 2014) that suggested such features may be realized by young educated speakers according to an implicational scale, and argued for variability and change as part of the complex system of speech. As for Spanish/English contact situations, evidence from elsewhere in the Southeast suggests that we might find varying patterns of dialectal accommodation among the Hispanic community (Wolfram et al. 2004). Analysis of vowel formants has not yet been conducted with Roswell Hispanic participants, a tantalizing possibility that deserves to be followed up. The partnership between the CVB and UGA has thus undoubtedly led to the kind of academic research that benefits the researchers and the university, and justifies their cooperation.

6 Living Laboratories


Meanwhile, a new possibility for the partnership emerged in Europe. The European Network of Living Laboratories (ENoLL) is intended to be a user-driven innovation system centered on ICT (information and communication technology), which is to say that customers or clients are involved in research and development from the earliest stage, up through the early market stage of products.⁶ Also important are SMEs (small to

medium enterprises), which are assisted by Living Labs in collaborative planning and organization. Large companies like Philips and SAP are also involved. Living Labs cross three domains: industrial (manufacturing, product development), social (change user habits), and territorial (regional development, 'open society', well-being, tourism). The *Roswell Voices* partnership fits into the last of these very well. *Roswell Voices* offers a model for understanding how communities create themselves, a new spin that has resonated with a number of Living Lab directors. The *Roswell Voices* Living Laboratory is the first, and so far still the only American member of the network (see Fig. 6.2); there are now half a dozen Living Labs in Canada, 4 in Mexico, and a total of 340 Living Labs at this writing in 50 countries worldwide. There is a serious review process for new members, using teams of current members and a five-point evaluation scale; 83 of 116 applicants (around 70 per cent) were accepted in the 'fourth wave' in which *Roswell Voices* became a member. I had been encouraged to apply by colleagues in Finland with whom I had been collaborating on humanities computing, and who thought that our work in Roswell offered good ideas for their activities in their Northern Urban/Rural Living Laboratory (NorthRULL).

The European Union expects to benefit from Living Labs through economic growth driven by innovation. The commercial side of Living Labs is meant to lower risk of investment by business. It creates a movement between developers and clients/customers of research, which constitutes a demand side policy. The network offers contacts in Living Labs when SMEs change regions or countries in Europe. ENoLL also participates in thinking globally and so encourages Living Lab members all around the world, but also in linking back to local affairs in Europe, which has been called a 'glocal' approach. ENoLL encourages local cooperation to compete globally, with global partners to inform local affairs, and it expects immigration to be a driver of globalization. Finally, ENoLL seeks to mobilize people for technology transfer. The relevance of all this for *Roswell Voices* is that ENoLL promotes the integration of local 'memory' (as a sense of place, of history) with business. Local communities need

⁶ For further details, see the ENoLL website, <http://www.enoll.org>.

Living Labs



Map Satellite

Google

Terms of Use Report a map error

news about us events all living labs members

ENoLL

Login or Sign up!

Roswell Voices LL

Roswell Voices first establishes, through language and life interviews, what communities exist within greater Roswell, such as historic families and houses around the Square, the historic African American neighborhood and churches, its new residential patterns and school districts, and its new Latin residents. The Roswell Convention and Visitors Bureau (CVB) can and does use the information to promote "Historic Roswell" both within the community and for visitors. Business and government, as time goes on, can use our information to improve delivery of goods and services.

Description

The Roswell Voices project began in 2002 as a partnership between the Roswell Convention and Visitors Bureau (in Roswell, GA, USA) and academic researchers at the University of Georgia (UGA; in Athens, GA, about 60 miles distant). Such a partnership is highly unusual in America, but we think it fits exactly what ENoLL is designed to create. Roswell Voices is an umbrella under which community development efforts, both cultural and business-oriented, can be associated with the historical and contemporary language and life of the community. We are committed to the use of emerging technology in information science and communications for community goals, in particular the use of university technical resources and experience: 1) to benefit visitors and residents of Roswell both personally and in commercial settings, 2) to benefit university students (in the American "service learning" movement), and 3) to benefit academic research (in advanced study of emerging network patterns in "real world" Roswell).

Information

- Information doc: Roswell_Voices.pdf
- Contact: Professor Bill Kretzschmar Department of English, University of Georgia, Athens, GA 30602 Email: kretzsch@uga.edu
- Website: <http://www.uga.edu/>
- Country: USA

Fig. 6.2 Roswell Voices on the ENoLL website⁷

to know who they are in order to perform well on the global stage. The great fear of globalization is the loss of local culture, whether in the third world or in the European first world, as global business trends threaten to swamp any sense of local identity. As a French delegate explained to

⁷ See <http://www.enoll.org> and <http://www.enoll.org/ourlabs/USA>.

me, it is all about 'la mémoire' in the face of conversion of historical communities by yuppification or gentrification or commercial change. *Roswell Voices* has established what Roswell, Georgia, has been and what it is becoming now in the voices of its residents, and the same could be done for communities across Europe and the world.

Living Labs are meant to be public/private partnerships (PPPs). Some are controlled by businesses like Philips, and others emerge from universities or NGOs. Some are devoted to product testing, and some to discussion of ideas like community risk management. Some have substantial investment in buildings and staff, some are 'virtual' like the *Roswell Voices* partnership. All are supposed to be long-term, not just focused on a single product. The central personnel tend to be true believers in principles of trust, openness, and bottom-up humancentric collaboration. The way that they talk about it, this is all intended to make an innovation ecosystem for user-driven cocreative research and development. In other words, ENoLL is really a movement for networking and best practices, not just about making more money. The ENoLL Council is managed by a Chair (currently Jarmo Eskelinen from Helsinki), and the ENoLL offices are in Brussels. Still, at a meeting I attended in Valencia at which new members including *Roswell Voices* were announced, speakers included the Head of the European Union Innovation Directorate, and two ranking members of the European Commission. There have been increasing financial commitments during each EU Presidency (starting with Finland) since ENoLL started in 2006. While ENoLL per se does not fund projects, its activities have been strongly supported by the European Commission. Funded initiatives have included:

- Energy Saving (Smart monitoring in Living Labs, data stored remotely for several small cities, with demos in public buildings and larger programs in school systems and public housing. The initiative sought to change social and behavioral patterns.)
- Healthcare (Mostly public health rather than direct clinical involvement, like ICT for home healthcare for the aging but some innovations like development of surgery robots were also supported.)
- eManufacturing, product development (ICT enhancements in the home)

The initiative just completed was about the Future Internet, which offered €600 million over three years. Collaborative funding was the pre-

ferred model, which meant PPPs in Europe with some possibility for joint local and EU funding for external members. The new EU Horizon 2020 funding scheme, which offers €80 billion over seven years, explicitly addresses the PPP model of cooperation between SMEs and universities.

The opportunity here for linguists is a new way to reach communities of speakers. While not all communities will be like Roswell, eager to engage with their linguistic status, ENoLL offers a chance to work with communities as part of a government-sanctioned program that embraces cultural studies. Linguistics can help communities develop their linguistic identity, as we have done in Roswell, both for the community's benefit in marketing itself or remembering its roots in the past, and for the linguists' benefit in the creation of a new and interesting data set. Linguistic interviews might be conducted as part of one of the funded ENoLL programs like healthcare or energy management, or they could take place on a cooperative shoestring as ours have in Roswell. There is no risk in a PPP other than the time and resources that linguists want to commit to the creation of a new resource. The Roswell CVB has not yet been interested in participating in a funded ENoLL program, but they have been very interested in the kind of outreach that ENoLL offers. The natural affinity of the PPP that we originally developed in Roswell with the Living Labs movement makes for a natural pathway of development.

7 The Future

The original idea for study and documentation of language and life in Roswell came from the example of Walt Wolfram's work in North Carolina. *Roswell Voices*, however, did not merely replicate that work; it extended it into a more formal partnership that can address more aspects of community life, including its economic patterns. As linguists, we would do well to assert ourselves as people who can engage with communities and help them to reach their full potential economically as well as culturally. If the ENoLL movement is right about the potential for bottom-up humancentric collaboration to make an innovation ecosystem with user-driven cocreative research and development, as many big words as those are, sociolinguists have a role to play in documentation of 'la mémoire' in many communities. We can help communities to find themselves, before they become lost

in globalization, and then to make glocal choices for their local livelihood. As linguists, we underplay our value if we only try to give something back to the community: we can help to make communities what they want to be in a global world. We plan to continue our involvement with Roswell, both with English and with Spanish speakers. In so doing, we bring value both to the community and to academic linguistics.

References

- Anderson, Bridget. 2005. The *cot/caught* merger in suburban Atlanta as a case of phonological leveling. Paper presented at NWAV 34, New York.
- Andres, Claire, and Rachel Votta. 2007. AAE and Anglo vowels in a suburb of Atlanta. Paper presented at LSA, Anaheim.
- . 2009. African American Vernacular English: vowel phonology in a Georgia community. In *African American English in Context, PADS 94*, eds. Malcah Yaeger-Dror, and Erik R. Thomas, 75–98. Durham: Duke University Press.
- Bailey, Guy, and Erik R. Thomas. 1998. Some aspects of AAVE phonology. In *African American English: Structure, History and Use*, eds. Salikoko S. Mufwene, John R. Rickford, Guy Bailey, and John Baugh, 85–109. London: Routledge.
- Childs, Becky, and Christine Mallinson. 2004. African American English in Appalachia: dialect accommodation and substrate influence. *English World Wide* 25: 27–50.
- Dunn, Joshua. 2008. Roswell voices: oral history and linguistics in Roswell, Georgia. Paper presented at CURO symposium, Athens.
- Garreau, Joel. 1991. *Edge City: Life on the New Frontier*. New York: Doubleday.
- Heath, Shirley Brice. 1983. *Ways with Words: Language, Life, and Work in Communities and Classrooms*. Cambridge: Cambridge University Press.
- Hilliard, Sarah, and Jeanine Carpenter. 2004. Vocalic alignment of Roanoke ‘Oisland’. Paper presented at the SECOL conference. Tuscaloosa: University of Alabama.
- Kretzschmar, William A. Jr. 2005. Language status and language change. Paper presented at SHEL-4, Flagstaff.
- . 2014. Complex systems in the history of American English. In *Developments in English: Expanding Electronic Evidence*, eds. Irma Taavitsainen, Merja Kjöto, Claudia Claridge, and Jeremy Smith, 251–264. Cambridge: Cambridge University Press.

- Kretzschmar, William A. Jr. 2015. African American Voices in Atlanta. In *The Oxford Handbook of African American Language*, ed. Sonja Lanehart, 219–235. Oxford: Oxford University Press.
- Kretzschmar, William A. Jr., Jean Anderson, Joan Beal, Bartek Plichta, Karen Corrigan, and Lisa Lena Opas-Hänninen. 2006a. Collaboration on corpora for regional and social analysis. *Journal of English Linguistics* 34: 172–205.
- Kretzschmar, William A. Jr., Claire Andres, Rachel Votta, and Sasha Johnson. 2006b. *Roswell Voices, Phase 2*. Roswell: Roswell Folk and Heritage Bureau.
- Kretzschmar, William A. Jr., Becky Childs, Bridget Anderson, and Sonja Lanehart. 2004. *Roswell Voices*. Roswell: Roswell Folk and Heritage Bureau.
- Kretzschmar, William A. Jr., and Joshua Dunn. 2010. Complex systems and sociolinguistics in Roswell. Paper presented at NWAV 39, San Antonio.
- Kretzschmar, William A. Jr., Joshua Dunn, and Mi Ran Kim. 2011. Implicational scaling in southern speech features. Paper presented at ADS/LSA, Pittsburgh.
- Kretzschmar, William A. Jr., Sonja Lanehart, Bridget Anderson, and Becky Childs. 2007. The relevance of community language studies to HEL: The view from Roswell. In *Managing Chaos: Strategies for Identifying Change in English, Studies in the History of the English Language 3*, eds. Christopher Cain, and Geoffrey Russom, 173–186. Berlin: Mouton de Gruyter.
- McNair, Lisa. 2005. *Mill Villagers and Farmers: Dialect and Economics in a Small Southern Town*. In *PADS 90*. Durham: Duke University Press.
- Pederson, Lee. 1986–1992. *Linguistic Atlas of the Gulf States*. 7 volumes. Athens: University of Georgia Press.
- Rowe, Ryan, and Tyler Kendall. 2004. *Regional and social diversity in the development of rural southern AAE: The case of Princeville*. Paper presented at SECOL. Tuscaloosa: University of Alabama.
- Wilson, Anna. 2013. *Roswell Voices, Phase 3*. Roswell: Roswell Folk and Heritage Bureau.
- Wolfram, Walt. 1993. Ethical considerations in language awareness programs. *Issues in Applied Linguistics* 4: 227.
- Wolfram, Walt, Phillip Carter, and Becky Moriello. 2004. Emerging Hispanic English: New dialect formation in the American South. *Journal of Sociolinguistics* 8: 338–358.
- Wolfram, W., J. Reaser, and C. Vaughan. 2008. Operationalizing linguistic gratuity: From principle to practice. *Language and Linguistic Compass* 10: 1–26.
- Wolfram, Walt, and Erik R. Thomas. 2002. *The Development of African American English*. Oxford: Blackwell.
- Zelinsky, Wilbur. 1992. *The Cultural Geography of the United States*, Rev edn. Englewood Cliffs, NJ: Prentice-Hall.

7

The Diachronic Electronic Corpus of Tyneside English and The Talk of the Toon: Issues in Preservation and Public Engagement

Adam Mearns, Karen P. Corrigan,
and Isabelle Buchstaller

1 Introduction

The continuing expansion in the number of digitized corpora of natural language raises various questions about the best ways of presenting, promoting, preserving and future-proofing such resources for current and future users. To deal with such concerns of sustainability and dissemination, it may formerly have been considered more than sufficient simply to deposit a data set, ideally encoded in an appropriate standard format, with a digital repository such as the University of Oxford Text Archive (OTA),¹ thereby making it available to other interested scholars and students in higher education. Addressing these issues in the context of more recent corpus projects requires consideration of a greater

¹ See <http://ota.ox.ac.uk>.

A. Mearns (✉) • K.P. Corrigan
Newcastle University, Newcastle upon Tyne, UK

I. Buchstaller
Universität Leipzig, Leipzig, Germany

range of factors. In large part, this can naturally be linked to the growing emphasis placed on public engagement and impact by organizations in the higher education sector, including funding councils. There is now a greater impetus for researchers to pay more attention to, and cultivate a better understanding of, prospective users beyond the narrow academic environment, for example those in schools and the heritage sector, as well as the general public. This new orientation is evidenced for example by the Public Engagement with Research Strategy promoted by Research Councils UK (RCUK) and the establishment in 2008 of the National Co-ordinating Centre for Public Engagement (NCCPE).² Developing effective methods of engaging with such user groups in a manner that generates impact has a particular relevance in the context of sociolinguistic projects that have assembled data sets consisting of interviews with members of these groups or with others in the same speech community. Wolfram's notion of linguistic gratuity advocates reciprocity as a core principle that should underpin community-based studies (Wolfram 1993: 227; see also Wolfram et al. 2008).

Embracing this principle encourages us to move beyond the idea that public engagement is essentially a matter of presenting the results of research in an accessible and appealing way to a non-specialist audience. To be genuinely reciprocal, a community-based study should strive to respond, where practicable, to the particular expectations, interests and needs of the population concerned, even (and perhaps especially) where these are not exactly the same as the goals of the academic research. In doing so, scholars will inevitably need to deal with the fact that the community is not a single homogeneous mass. As the reference above to schools, the heritage sector and the general public indicates, establishing relationships with people across the community involves working in a variety of contexts, with groups whose expectations, interests and needs are potentially quite diverse. Beyond the idealistic reasons for pursuing constructive relationships of this kind, in the context of UK higher education there are also the practical incentives already noted.³ In this chapter,

² For further details, see the websites of RCUK (<http://www.rcuk.ac.uk>) and the NCCPE (<http://www.publicengagement.ac.uk>).

³ See, for example, the definition of impact provided on the RCUK website, 'Pathways to Impact' (<http://www.rcuk.ac.uk/innovation/impacts>).

we discuss how the issues of corpus sustainability and dissemination have been addressed in the creation of the *Diachronic Electronic Corpus of Tyneside English* (DECTE, Corrigan et al. 2012) and, in particular, in the design and development of the project's public-facing website, *The Talk of the Toon*.⁴ This open access site is at the heart of the public engagement activities through which the DECTE project seeks to fulfil the aim of fostering mutually beneficial and productive relationships with diverse non-academic user groups in the wider community. This stance is in keeping with the principles connected to research for empowerment already noted and the requirements of the current public engagement and impact agendas in UK higher education.

Sect. 2 below describes the current composition of DECTE, and the processes by which it has been constructed as a major update of the *Newcastle Electronic Corpus of Tyneside English* (NECTE, Corrigan et al. 2005; see also Allen et al. 2007). The main development in terms of content is that the legacy collections of sociolinguistic interviews from the 1960s–1970s and 1990s which were digitized, transcribed and combined to form NECTE have now been augmented with materials from the more recent monitor corpus, NECTE2, which was established in 2007.⁵ The addition of this newer component has greatly increased not only the size of the corpus, but also its geographic reach, which is spreading beyond the core Tyneside region covered by NECTE, to include speakers from other areas of the North East of England. In conjunction with this geographic growth, there is also a continuing expansion in the kinds of informants being drawn from the local communities that the corpus now covers, with participants representing a wider range of demographic groups than were captured by the earlier, more narrowly targeted phases of data collection. This increase in the number and type of informants is matched by a similar rise in the number of fieldworkers involved in the process. The reason for this latter development is that the annual collection of NECTE2 interviews is conducted

⁴Between October 2010 and January 2012, DECTE was funded by the Arts & Humanities Research Council (Grant: AH/H037691/1). *Toon* is a local term for Newcastle whose football supporters, for example, are commonly known as the *Toon Army*.

⁵See <http://research.ncl.ac.uk/necte2> and the documentation section of the DECTE website (<http://research.ncl.ac.uk/decte/documentation.htm>).

by undergraduate and postgraduate students at Newcastle University, as part of an ongoing teaching and learning initiative. The main focus of this chapter is on the ways in which the DECTE project engages with user groups outside higher education. This aspect of the current phase of corpus construction demonstrates that it is also very much concerned with employing methods of research-led teaching. The objectives are to cultivate the interest of sociolinguistics students at Newcastle, and in particular to deepen their knowledge and experience of fieldwork techniques, through involving them in an active project. Some of the specific issues that arise from the participation of a large number of student interviewers in the collection process will be explored in a little more detail in Sect. 2.

We will also discuss one further aspect of the corpus construction process, the importance of which has been highlighted by the current phase of DECTE's development. Given its status as a monitor corpus, with new sets of interviews collected annually as stated above, the integration of NECTE2 as a subcomponent of DECTE has turned the combined data set into an open-ended enterprise. Consequently, the question of how best practices for the longer-term preservation and sustainability of the resource can be guaranteed has become a particularly significant consideration. The discussion below will demonstrate our belief that the foundation for being able to address this issue effectively was laid during the earlier NECTE project, through the adoption of an underlying corpus architecture that conforms to the extensively used guidelines and standards established by the Text Encoding Initiative (TEI) for the digital representation of documents in Extensible Markup Language (XML).⁶ With respect not only to preservation, but also to the overall and continuing usefulness of the resource, we will suggest that the choice of TEI-conformant XML as the format for the corpus document files has a number of important advantages. Broadly speaking, these relate to the straightforward way in which it can be updated and maintained—a characteristic that is related to its status as a widely adopted and evolving

⁶For more information on the Text Encoding Initiative, see the organization's website (<http://www.tei-c.org>). For a description of the basics of XML, see for example the tutorial on the W3Schools website (<http://www.w3schools.com/xml>).

standard—and to the fact that it is a format that promotes compatibility between resources. This compatibility can manifest itself in various ways, for example in the ease with which a TEI XML corpus can be used with a range of text analysis software, or integrated with other similar data sets.

In Sect. 3, we will show how this benefit of DECTE's standardized architecture has also been essential to the creation of the previously mentioned public website, *The Talk of the Toon*, and therefore to the project's programme of public engagement, which is built around it. This section will also describe in detail how multimedia content has been organized according to a number of designated themes, relating to various different areas of the speakers' lives.

More generally, we believe that our public presentation of the corpus has also been successful in encouraging users from the North East of England to value their own linguistic identity as a product of their unique heritage, thereby embodying the principle of empowerment advocated by Cameron et al. (1997).

2 DECTE: Constructing and Preserving the Corpus

DECTE has been constructed by combining three separate data sets of sociolinguistic interviews. These components consist of audio recordings, transcriptions and associated materials such as social data files that were collected for three different research projects, spanning the last six decades. Fig. 7.1 summarizes the current structure of DECTE and the chronological phases of development.

2.1 The TLS and PVC Projects

As Fig. 7.1 shows, the earliest of the three projects that comprise DECTE was the *Tyneside Linguistic Survey* (TLS—Strang 1968; Pellowe et al. 1972; Pellowe and Jones 1978; Jones-Sargent 1983). This project began in the late 1960s and was established to investigate the nature and extent of variation in Tyneside speech, with the aim of discovering whether such

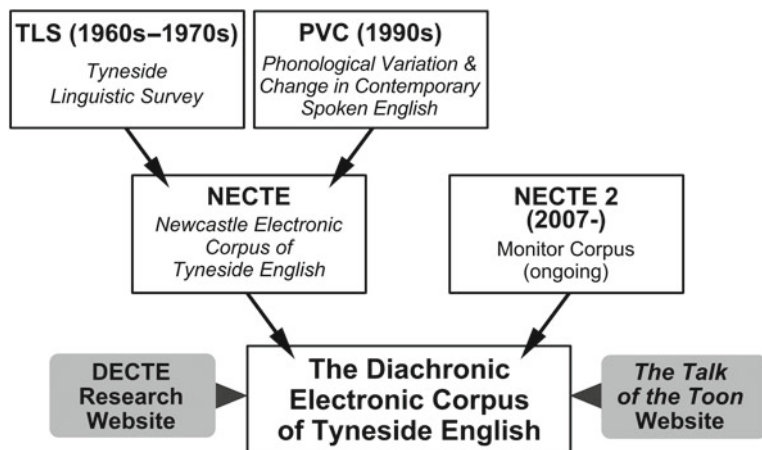


Fig. 7.1 DECTE: components and phases of development

variation correlated in any way with the social attributes of the speakers (Pellowe et al. 1972: 1). The initial focus was on Newcastle. The TLS project team proposed to record more than 250 speakers from the city in two phases, the majority being randomly selected from the Electoral Register, but with smaller subsets being individually chosen (Pellowe et al. 1972: 22–3). A third phase was also planned, with the aim of extending the survey across the River Tyne to Gateshead and interviewing a further 150 speakers, again selected randomly from the Electoral Register, though this time stratified by the rateable value of homes in different polling districts (Pellowe et al. 1972: 23–4). Data collection took the form of one-on-one interviews recorded in the participants' homes. Participants were given a list of words to read, in order to elicit particular accent features, and were also asked for their judgements on selected grammatical constructions and whether they used items of vocabulary that might be considered characteristic of the local dialect. However, for the most part, the recordings consist of normal conversations, albeit somewhat structured and sometimes slightly formal in style, given the one-on-one interview context and the lack of any prior acquaintance between the interviewer and the participants. Nevertheless, the interviewees were encouraged to talk about various aspects of their lives during these conversations,

including their family backgrounds and the jobs they had undertaken since leaving school, as well as their opinions on issues such as education and politics. As a result, many of the recordings contain precisely the kind of oral narratives of personal experience that allow the collection to serve not only as a data sets for linguistic analyses, but also as the basis for an oral history archive (see Labov 2004, 2006), as we shall see in the account of *The Talk of the Toon* website presented below.

Like the TLS, the *Phonological Variation and Change in Contemporary Spoken British English* project (PVC—Milroy et al. 1997; 1999; Watt and Milroy 1999; Watt 2002), which ran from 1994 to 1997, was designed to investigate correlations between accent features and social characteristics. Interviews for this project were conducted in both Newcastle and Derby. The intention was to analyse phonological variables that were undergoing change in non-standard varieties of British English, and to investigate the social trajectories of those changes ‘with particular reference to the spread of localized change to supra-local domains’ (Milroy et al. 1999: 37). In Newcastle, 18 interviews were recorded in 1994, each lasting around 60 minutes. The informants were recruited using a social network model. The final sample contained roughly equivalent numbers of males and females in two age groups (16–20s and over 40s) and from two neighbouring residential estates about four miles north-west of the city centre, one designated middle class (Chapel House) and the other working class (Newbiggin Hall) based on the judgements of the local fieldworker and on data from the 1991 Census (Milroy et al. 1999: 37; Watt 2002: 46). The participants were asked to read a list of around 200 words in order to prompt their use of particular speech sounds (Watt 2002: 46), but—as with the TLS interviews—the recordings mainly consist of normal conversations. Indeed, for the most part, these have a rather more natural, free-flowing quality than some of those of the earlier project, since they involve pairs of friends or relatives who speak to each other on a range of subjects with little or no intervention from the fieldworker. In this sense, they are perhaps even better suited to the dual purpose of linguistic data set and oral history archive, which they now serve as a sub-component of both DECTE and *The Talk of the Toon*.

2.2 NECTE: Combining the TLS and PVC Subcorpora

Between 2001 and 2005, the materials from the TLS and PVC projects were enhanced and amalgamated to form the *Newcastle Electronic Corpus of Tyneside English* (NECTE, Corrigan et al. 2005), a single TEI-conformant XML-encoded corpus that presents the interviews in a variety of time-aligned forms, linking the digitized audio not only with the corresponding orthographic transcription, but also with a part-of-speech tagged version of the text and, in the case of the TLS files, partial phonetic transcriptions.⁷ As far as the PVC subcorpus was concerned, the process of incorporating the material into NECTE was a fairly straightforward one, involving the production of full orthographic and part-of-speech tagged transcriptions of the 18 interviews, which had not been required for the purposes of the original project. When it came to the TLS subcorpus, the process was rather more complicated. The only surviving material that was discovered for the Newcastle part of this survey consisted of social data files and partial phonetic transcriptions associated with seven of the interviews. In contrast, far more of the material linked to the project's planned third phase of 150 interviews in Gateshead was recovered. Reel-to-reel tapes, social data files and partial phonetic and orthographic transcriptions relating to 107 interviews were identified, though not all had full sets of each of these components. Moreover, some of the audio recordings were damaged or had deteriorated due to the passage of time so while the orthographic transcriptions remained, in some cases there was no audio to link them with.⁸ After digitizing those analogue recordings that were salvageable, the NECTE project team was able to create full orthographic transcriptions for 88 of the Gateshead interviews. The average duration of these recordings is around 40 minutes and, judging by references made in the interviews for example to contemporary events or the age and birth dates of the informants, they were all apparently

⁷ For a full account of the creation of NECTE, see Allen et al. (2007) and the documentation pages on the NECTE (<http://research.ncl.ac.uk/necte/documentation.htm>) and DECTE (<http://research.ncl.ac.uk/decte/documentation.htm>) websites.

⁸ In December 2013, an additional 23 previously unknown reel-to-reel tapes were discovered. These appear to contain a further 13 Gateshead recordings and interviews with perhaps as many as 60 Newcastle informants. It is not yet clear how many of these interviews will be recoverable.

recorded in 1971–72. Of this group of 88, a total of 37 had complete sets of the various interview components outlined above. These 37 TLS interviews were incorporated into the corpus proper—that is, the set of TEI-conformant XML-encoded NECTE files—together with the 18 PVC interviews (Allen et al. 2007): 21–35). Despite the absence of the corresponding audio recordings, XML files of the seven partial phonetic transcriptions of the Newcastle TLS speakers were also included in the corpus, because of the valuable opportunity they nevertheless afford for comparing the characteristics of contemporary speakers from the two neighbouring areas of Newcastle and Gateshead.

2.3 NECTE2: The Monitor Corpus

DECTE has been created by augmenting the legacy materials contained in NECTE with NECTE2, a new and ongoing subcorpus established at Newcastle University in 2007. This subcomponent consists of interviews primarily collected by undergraduate and postgraduate students of English language and linguistics who are taking modules in sociolinguistics.⁹ For one of the assignments that form the assessment of their course, the students are tasked with recruiting two informants who must have lived for 95 per cent of their lives in the area covered by the local government regions that make up the North East of England. These are Tyneside (the metropolitan boroughs of Newcastle, Gateshead, North Tyneside and South Tyneside), Northumberland, County Durham, Wearside (the metropolitan borough of the City of Sunderland) and Teesside (the Boroughs of Hartlepool, Darlington, Stockton-on-Tees, Middlesbrough and Redcar and Cleveland).¹⁰ The wider geographic

⁹The first phase of turning the new collection of interviews into a corpus was funded in 2009–10 by an award for research collaboration and infrastructure initiatives made to Isabelle Buchstaller, Karen Corrigan, Gerard Docherty and Ghada Khattab by Newcastle University's Centre for Research in Language and Linguistics Sciences (CRiLLS, <http://www.ncl.ac.uk/linguistics>). As well as members of the DECTE project team, the modules associated with the NECTE2 data collection have at various times been taught by Lynn Clark, Heike Pichler, Jennifer Thorburn and Cathleen Waters.

¹⁰For a map, see the documentation section of the DECTE website (<http://research.ncl.ac.uk/decte/documentation.htm>).

reach of this latest phase of the corpus gives the student interviewers a much larger pool of potential informants, making recruitment somewhat easier. It also has the added advantage of reducing the 'observer's paradox' because interviews are more likely to be conducted with friends and family (Labov 1972; Milroy 1987). More importantly, however, this wider regional coverage benefits the corpus in making the data set more broadly representative of diverse local north-eastern communities. Not only does this approach benefit subsequent linguistic analyses, but it also increases the potential appeal of the interviews in the website context.

Once the students have enlisted suitable volunteers and collected detailed information about their social background, they record an informal interview with their participants, lasting around 60 minutes, and transcribe a 30-minute section of this interview. They also make recordings of their informants reading word lists and a short passage of text. In many ways, then, the main interviews that the students record reflect the same sort of approach that was taken with the PVC data collection of the 1990s: they involve dyadic pairs of informants who are typically friends or relatives and therefore often quite similar in terms of age or social background, or both. The familiarity that the informants have with each other (and often also with the interviewer) means that these conversations frequently have the same free-flowing quality and wide-ranging coverage of topics that is found in the PVC recordings, and are therefore equally well suited to being part of the oral history component of *The Talk of the Toon* archive.

One area where the NECTE2 interviews differ somewhat from those of the PVC subcorpus is in the way that the fieldworkers perform the role of interviewer. In almost all cases, the student interviewers are more active than the PVC fieldworker was in directing the conversation using the prompts supplied by an interview protocol which they have devised themselves to match their informants' interests. However, the extent to which they utilize this tool to best effect, or indeed join in as full participants in the conversation, is highly variable across the subcorpus. Although students are given comprehensive advice on how they can approach the recordings in order to get the best out of their informants, the degree of variability is understandable, especially given the fact that in the seven academic years from 2007–8 to 2013–14, 482 interviews have been added

to the NECTE2 monitor corpus, meaning that roughly the same number of different interviewers have been involved in the process.¹¹

In other aspects of the NECTE2 data collection process, measures are taken to constrain as much as is practicable the variability that can arise from the involvement of so many different fieldworkers. For example, students are given detailed templates for all of the documents they will need to complete their assignment. These include Demographic Text Files to record all of the required social background details of their informants and a Transcription File in which to transcribe the 30-minute segment they select from their recording (for an example of the latter, see the extract in Fig. 7.2). The Transcription File is simply a regular Microsoft Word document. The main reason that students are currently instructed to transcribe their interviews in Word, rather than using dedicated transcription software such as ELAN,¹² is that the transcription assignment is only one part of the assessment and content of the sociolinguistics module to which it is presently attached. We have found that using a program with which students are already familiar, and keeping the technical aspects of the task at a relatively modest level, allows a suitable balance to be struck between fitting the preparation for the assignment into the course while also being able to focus on the broader insights that it offers about sociolinguistic data collection and fieldwork techniques, as well as about some of the issues of language variation that inform the module as a whole. The focus on the latter has been heightened by requiring students to identify and annotate variants of two selected linguistic variables, such as (ing) and h-dropping, in the speech of their informants. They are also naturally assessed on the accuracy of their transcriptions as representations of the audio recordings, and on the extent to which they adhere to the conventions of the prescribed transcription for-

¹¹The number of interviewers is not exactly the same as the number of interviews because, from 2008 to 2010, students taking a course in language variation and change in the British Isles collected interviews as part of a group assignment, and in some instances more than one member of the group took part in the recording. In a very small number of cases, students have also been involved in the data collection more than once, by virtue of taking more than one of the modules that has been linked to the monitor corpus since its inception.

¹²ELAN is a program for adding annotations to audio and video files, created at the Max Planck Institute for Psycholinguistics (The Language Archive, Nijmegen; see <https://tla.mpi.nl/tools/tla-tools/elan>).

3. Conventions for Transcribing Oral Discourse Phenomena

3.1. False Starts

If a speaker utters only part of a word, indicate that it is incomplete with a hyphen (-) at the end of the word fragment, e.g.

[2014/IN/PI/1234] do you ha- do you have any sort of particular attachments to Tyneside

There is no space between the word fragment and the hyphen; there is a space after the hyphen.

Hyphens can also be used in normal hyphenated constructions (e.g. semi-skimmed, A-Levels) and in some fillers (see §3.6 below). They are not used in any other contexts.

Correction Pass: 2

Transcription start time: 5 minutes 46 seconds

SPEAKER	UTTERANCE
[2013/KS/8098]	I prefer an Indian a really hot curry
[2013/AK/8098]	I can't imagine you like handling it
[2013/KS/8098]	too hot that I my face goes red
[2013/AK/8098]	<@> and your nose runs
[2013/KS/8098]	yeah and I'm like yes <@>
[2013/AK/8098]	I always have curries in the winter
[2013/IN/MS/8098]	to warm you up
[2013/AK/8098]	yeah
[2013/KS/8098]	yeah
[2013/AK/8098]	and to get rid of cold

Fig. 7.2 Extracts from the NECTE2 Orthographic Transcription Protocol (*top*) and a completed NECTE2 Transcription File (*bottom*)

mat. For guidance on formatting, students are provided with a detailed Orthographic Transcription Protocol (OTP), setting out the appropriate notations for representing various features of the participants' conversation. This is based on the OTP developed for the NECTE project (see Allen et al. 2007): 22–4; Beal et al. 2014: 519–23), adapted to suit the requirements of the current assessment context and evolving over time as we have monitored how successive cohorts of students have dealt with the demands of the task. Fig. 7.2 shows a section from the latest version of the NECTE2 OTP, illustrating the conventions to be used when representing false starts, and an extract from a completed NECTE2 Transcription File submitted by one of the students.

We have made reference at various points above to the fact that the documents incorporated in DECTE, like those in NECTE before it, take the form of a set of XML files which are encoded according to the widely

used guidelines and standards set out by the TEI. In Sect. 1, we also briefly alluded to some of the advantages of the TEI XML format, noting that it promotes compatibility and is easily updated. The latter is crucial in helping to ensure the long-term sustainability of a digital resource, given the speed with which standards and formats inevitably evolve, in line with rapid developments in the field of Information Technology in general. Indeed, this is an issue that we have already had to tackle, since part of the process of developing DECTE as an update of the earlier NECTE resource involved revising the existing files in order to make them comply with the latest version of the TEI guidelines, which was introduced in 2007, two years after the release of NECTE.¹³ The quality of compatibility facilitated by the TEI XML format is a beneficial feature for a number of reasons. First and foremost, because it results in data sets that are independent of the specific characteristics of individual computer platforms and are not produced in and therefore tied to one particular software package, it means that they can be used relatively effortlessly with a range of XML-aware products, including various programs that are designed for textual analysis, thus enabling users to work with the corpus in those applications that best suit their needs or preferences.¹⁴ A related benefit is that, as a widely adopted specification, the compatibility afforded by the TEI XML format fosters interoperability with an extensive range of other data sets and resources that are also built around this standard, as for example in the ENROLLER portal, which integrates a diverse set of electronic resources—including NECTE/DECTE, as well as others, such as the *Scottish Corpus of Texts & Speech* (SCOTS) (see Anderson and Hough, this volume)—into a single online repository, thereby allowing for cross-searching of the combined collections.¹⁵ In Sect. 3, we will return to a further benefit that relates to the compatibility that stems from DECTE's standardized architecture, namely the

¹³ For full details of this latest version of the guidelines, designated 'P5', see the TEI website (<http://www.tei-c.org/Guidelines/P5>).

¹⁴ See, for example, *TXM* (<http://textometrie.ens-lyon.fr>), Mike Scott's *Wordsmith* (<http://www.lexically.net/wordsmith>) and *Xaira* (<http://xaira.sourceforge.net>).

¹⁵ For more details, see the ENROLLER (Enhanced Repository for Language and Literature Researchers) portal (<https://enroller.nesc.gla.ac.uk>).

ease with which the corpus interview files could be incorporated into the *The Talk of the Toon* website.

A detailed description of the TEI-conformant XML format of the corpus files can be found on the DECTE website,¹⁶ with further accounts presented by Allen et al. (2007: 33–5) and Beal et al. (2014: 524–32). For the purposes of the current discussion, it is sufficient simply to note some of its basic characteristics. Essentially, XML involves marking particular elements or sections of text within a document, by enclosing them in tags that define their category or function. In principle, these tags can be deployed to label any aspect of the text, using any descriptive term that the person creating the document might consider appropriate. The TEI guidelines constitute a standard in the sense that they prescribe sets of necessary and optional tags, and the form that these can take, thereby defining the overall structure and configuration of a well-formed XML document. As well as general conventions that must be followed in all cases, the guidelines stipulate others that apply to specific kinds of material, in order to address the particular requirements of a diverse range of text types, such as correspondence (see Amador-Moreno et al., this volume), manuscripts, performance texts, verse and—as in the case of DECTE—transcriptions of natural speech. The extract in (1) below is a section of `decten2y10i009.xml`, one of the TEI XML interview transcription files from the NECTE2 subcorpus of DECTE, illustrating for example the way in which each utterance is demarcated with an opening tag, such as `<u who="#informantY10i009a">`, and ends with a closing `</u>` tag. The ‘Y10i009a’ part of the opening tag identifies the speaker in question using a unique identifier code and therefore serves to associate the utterance with that informant. Other elements include tags that designate non-linguistic incidents, such as `<incident><desc>interruption</desc></incident>`, which denotes the point in an utterance where another participant begins to speak, therefore overlapping with the current speaker, and time anchor tags, such as `<anchor xml:id="decten2y10i009ortho0060"/>`, which functions to align the orthographic transcription to the audio by marking those points in the text corresponding to

¹⁶ See, for example, the pages on ‘Corpus Structure’ in the Documentation section (<http://research.ncl.ac.uk/decte/structuring.htm>).

20-second intervals in the associated sound file (in this case '0060' signals the 1 minute point).

(1) **An extract from decten2y10i009.xml**

```
<u who="#informantY10i009b"> I – I think
<incident><desc>interruption</desc></incident> th- th- they're
just as good I – I mean yo- yo- you're – you're talking about like
de- distinguishing <anchor xml:id="decten2y10i009ortho0060"/>
between the Mackems </u>
<u who="#informantY10i009a"> Yeah I'm talking about in foot-
ball terms. </u>
<u who="#informantY10i009c"> Ah </u>
<u who="#informantY10i009b"> In football
<incident><desc>interruption</desc></incident> terms </u>
<u who="#informantY10i009a"> <incident><desc>interruption
</desc></incident> In football terms er we're – they're at each
other's throats </u>
<u who="#informantY10i009b"> Well just find that </u>
<u who="#informantY10i009a"> But in if you were to go over
and talk to them normally if they didn't know what you were
they'll talk to you </u>
```

The format of this document is clearly rather different from that of the student's NECTE2 Transcription File illustrated in Fig. 7.2. The process of turning the transcriptions submitted by students into the TEI XML files required for the main corpus is a labour-intensive one, involving a number of stages that need to be completed before the introduction of the kinds of tags seen in extract (1). In addition to the initial step of checking for and correcting any errors in the transcription, focusing on its accuracy both as a record of the audio and in its application of the OTP conventions, it is worth highlighting two further issues. Both are significant not only with respect to the general design of the corpus, but also specifically in terms of the presentation of the interviews as part of *The Talk of the Toon* website. The first has to do with establishing the alignment that links the interview transcriptions to the related sound files.

In keeping with the method adopted for the NECTE data set, achieving this currently requires the manual insertion of provisional markers throughout the length of the transcription, at points corresponding to 20-second intervals in the audio. Although these temporary markers can subsequently be very easily transformed into the TEI XML time anchors mentioned above, the process of inserting them is inevitably rather protracted. Nevertheless it is also worthwhile: the time alignment is an indispensable feature of the corpus and a vital factor in the development of the *Toon* website, with the anchor tags forming the basis for the functionality of the site's text/audio interface, as described in Sect. 3 below. The second major issue that must be addressed is a matter of content relating to the ethical management and presentation of the interview material. To protect the anonymity of participants and any private individuals they mention, the interviews need to be carefully checked so that personal names and other references that could identify people can be tagged for later substitution (in both the transcriptions and the audio files). The importance of this process of anonymization is of course emphasized in the context of *The Talk of the Toon*, which makes the interview materials freely accessible online. We will therefore return to this issue in Sect. 3 (see also Cheshire and Fox, this volume).

Given the time-consuming nature of this transformation process, only a subset of the 482 interviews collected between 2007–8 and 2013–14 have so far been fully incorporated into DECTE, with 44 being reformatted and added to the main XML-encoded corpus during the period from the beginning of the project in 2010 to the release of the data set at the beginning of 2012. There is also a group of TLS interviews that were not included in the earlier NECTE phase of the data set, because they did not have a complete set of all of the various components associated with that subcorpus. Table 7.1 summarizes the current composition of DECTE and its three constituent parts, indicating the size of both the core XML-encoded corpus and the larger collections associated with the TLS and NECTE2 subcomponents.

We conclude this section with details of one further set of files, which is available as a supplement to the main corpus. Although we have focused above on some of the main benefits that DECTE derives from employing the TEI XML file format, and will explore some further aspects of these in discussing the *The Talk of the Toon* website below, we have also taken the

Table 7.1 The current composition of DECTE

	DECTE	Components		
		TLS	PVC	NECTE2
Recording dates	1971–2013	1971–1972	1994	2007–2013
<i>XML-encoded corpus</i>				
Interviews	99	37 ^a	18	44
Words	804,266	229,909	208,295	366,062
Audio (hr:min:sec)	71:45:43	22:53:55	17:34:25	31:17:23
Informants ^b	160	37	35	88
Female	87	20	18	49
Male	73	17	17	39
<i>Full collections</i>				
Interviews	588	88 ^a	as above	482
Words	4,774,406	584,432		3,981,679
Audio (hr:min:sec)	404:54:25	59:43:38		328:36:22

^a The corpus also contains seven phonetic transcriptions of Newcastle informants. There are no orthographic transcriptions or audio recordings for these interviews, so they are not included here.

^b The PVC and NECTE2 interviews have two informants per interview, while the TLS has one. There are 35 (rather than 36) informants recorded for the 18 PVC interviews because one participant was recorded twice.

decision to create alternative, plain text versions of the processed interview transcriptions. While we continue to believe that the use of TEI XML is the best way of ensuring the longevity and compatibility of the data set, it has also become apparent to us that, for some users, the XML format is an obstacle that can impede them in working with the corpus, or deter them from using it altogether. The reason for this is that, despite the XML files working perfectly with software that can properly interpret the tags (which need never be seen by the user), they are not suited to being opened and read in the same simple way that a plain text file can be. In other words, although users can choose from a range of compatible XML-aware software, they do need some such program in order to access the files as intended. For those users who have no experience with any programs of this kind, and no incentive or inclination to develop that experience, the fact that the corpus interview files appear to be cluttered with complicated labels when viewed in an everyday text editor or word processor can therefore be a problem. The DECTE plain text files are thus intended to assist this category of prospective users, including for example researchers and university students who do not typically work with electronic corpora, or

teachers and pupils in schools and colleges. Extract (2) presents the same interview passage as was seen in extract (1), but this time in its plain text format. A comparison of the two clearly illustrates the fact that the plain text copy of the transcription is a much simplified version of the XML file. For example, the speaker labels that mark the beginning of each utterance are more basic in form, with the <Informant Y10i009a> that appears in the plain text file equating to the <u who="#informantY10i009a"> seen in the XML version above. This simplified format enhances readability, while containing enough metadata in the labelling to be able to represent the main characteristics of the content and structure of the interview. The fact that the labels share the same basic form as an XML tag, for example through the use of the angle brackets, also means that these plain text files can be used in a straightforward way with text analysis programs that can work with tagged texts but do not necessarily require the full XML configuration, such as *AntConc* (Anthony 2015).¹⁷

(2) An extract from decten2y10i009.txt

<Line 0023><Informant Y10i009c> <interruption> <unclear> and people from Jarrow
 <Line 0024><Informant Y10i009b> I – I think <interruption> th-th- they're just as good I – I mean yo- yo- you're – you're talking about like de- distinguishing <time: 1 min> between the Mackems
 <Line 0025><Informant Y10i009a> Yeah I'm talking about in football terms.
 <Line 0026><Informant Y10i009c> Ah
 <Line 0027><Informant Y10i009b> In football <interruption> terms
 <Line 0028><Informant Y10i009a> <interruption> In football terms er we're – they're at each other's throats
 <Line 0029><Informant Y10i009b> Well just find that
 <Line 0030><Informant Y10i009a> But in if you were to go over and talk to them normally if they didn't know what you were they'll talk to you
 <Line 0031><Interviewer> Yeah

¹⁷ See the *AntConc* website (<http://www.laurenceanthony.net/software/antconc>).

3 The Talk of the Toon: Presenting and Promoting the Corpus

We have already touched upon some of the features of the corpus structure and content that we have been able to take advantage of in designing the project's *Talk of the Toon* website. The interactive functionality of this multimedia site crucially depends upon and exploits DECTE's TEI-conformant XML file format. The multimedia content has been compiled by combining DECTE's interview transcriptions and audio recordings with pictures and video clips that similarly illustrate aspects of the recent history of the North East and its people. In this section, we focus on the process by which this public-facing website was created, the functionality it offers, and the range of public engagement activities that have both helped to develop and then utilized the material it contains.

As a means of broadening its appeal and potential applications, adopting this approach to the configuration of the material has been fundamental in pursuing the aim of establishing DECTE as an accessible public corpus that meets the needs of a range of general users (see Kretzschmar et al. 2006: 179–82). It will become apparent that involving prospective users—especially those in schools and colleges—in the design of this organizational structure, and in the overall presentation of the material, was key to its success. This strategy laid the foundation for *The Talk of the Toon* to be promoted effectively as a tool suitable for a range of curriculum subjects at all levels of education and indeed for continuing professional development purposes as Cheshire and Fox (this volume) argue. Fig. 7.3 shows a screenshot of the website's home page.

The links give a sense of the overall structure of the site. The Interview Index link redirects the user to a page that lists all of the available corpus interviews, indicating the time period in which they were recorded and giving a brief overview of the demographic details of the participants (their sex and age group). From here, the user can select one of the interviews and click through to its interface page, which provides access to the relevant transcription and audio recording. Fig. 7.4 contains an example of one of these interface pages, in this case that for the Y10i009 interview from the NECTE2 subcorpus which we have already seen in extracts (1) and (2).

The Talk of the Toon
An archive of local language and stories
 the memories, thoughts and opinions of the people of Tyneside, past and present, in their own words

Home | About | Interview Index | Themes | Quizzes | Schools | Top Stories | Introduction to North East Dialects | Links

Themes
 Family, Home, Work, Shopping, Sport, Entertainment, etc.
[\[Click Here\]](#)

Word Search
 Search the Archive
[\[Click Here\]](#)

Quizzes
 How well do you think you know the Geordie dialect?
[\[Click Here\]](#)

Schools
 Guidance for Key Stages 2 and 3, GCSE and A-Level
[\[Click Here\]](#)

The Talk of the Toon

- [About](#)
- [Feedback](#) | [Get Involved](#)
- [Introduction to North East Dialects](#)
- [Interview Index](#)
- [Links](#)
- [Privacy Policy: Cookies](#)

Top Stories — Our First Telly: channel hopping

1990s / Female 41-50 & Female 41-50

It was only one channel, BBC ... that was our first telly ... and then when Tyne Tees came out, I used to go down to the neighbours, two doors down, to watch the other programmes on the Tyne Tees ...

[Hear the full story](#) — [See more Top Stories](#)

Picture: Kevin Dean (2010, CC BY-NC-SA 2.0)

Home | About | Interview Index | Themes | Quizzes | Schools | Top Stories | Introduction to North East Dialects | Links

Contact: decte@ncl.ac.uk | [Feedback Form](#) | [Get Involved](#)

Fig. 7.3 The home page of *The Talk of the Toon*

On the left-hand side of the page are the audio player panel, with a volume slider and the usual sort of buttons for controlling playback, and the interview transcript panel containing the text of the interview. Each speaker's utterances appear in a different colour in the transcript. The colours are not only intended to enhance the overall appearance of the page, but more importantly also to provide a clear visual link with the speaker panels on the right-hand side. There is a panel for each of the participants, matching the colour of their utterances in the text. The panels that relate to the main participants (that is, the interviewees, rather than the interviewer or others who happened to be present at the time of the recording) summarize their background details, with information on their age group, gender, area of residence, education and occupation. This information is acquired

The Talk of the Toon

Home About Interview Index Themes Quizzes Schools Top Stories Introduction to North East Dialects Links

Archive Interview: Y10i009

Return to: [Theme Results](#) | [Interview Index](#)

For a guide to the layout of this interview page and how to use it, [click here](#).

decten2y10i009audio

Interview Transcript

They're *(interruption)* canny lads

Speaker 4: *(interruption)* I work with them

Speaker 2: Well I work with Mackems

Speaker 4: *(interruption)* *(unclear)* and people from Jarrow

Speaker 3: I -- I think *(interruption)* th- th- they're just as good I -- I mean yo-yo- you're -- you're talking about like de- distinguishing between the Mackems

Speaker 2: Yeah I'm talking about in football terms.

Speaker 4: Ah

Speaker 3: In football *(interruption)* terms

Speaker 2: *(interruption)* In football terms er we're -- they're at each other's throats

Speaker 3: Well just find that

Speaker 2: But in if you were to go over and talk to them normally if they didn't know what you were they'll talk to you

Speaker 1: Yeah

Speaker 2: And we'll -- and they'll talk to us I mean

Speaker 1: So is it -- because I'm aware of it but I didn't -- is it a massive rivalry or?

Speaker 1: interviewerY10i009

Speaker 2: informantY10i009a
Age Group: 41-50
Gender: Male
Residence: Tyneside - Newcastle (born in North Shields, North Tyneside)
Education: Left school at 16
Occupation: Driver

Speaker 3: informantY10i009b
Age Group: 51-60
Gender: Male
Residence: North Tyneside - North Shields
Education: Left school at 16
Occupation: Joiner

Speaker 4: informantY10i009c

Speaker 5: informantY10i009d

Themes

Click a theme in the menu below to highlight related keywords in the transcript.

- Language & Identity
- Community, Politics, Law & Order
- Work & Industry
- Sport
- Nightlife
- Transport
- Entertainment & Culture
- Clothes & Fashion
- Home
- Money & Shopping

Fig. 7.4 The interview interface page for decten2y10i009

from the metadata recorded at the beginning of each interview's XML file. Indeed, the transcription text itself is also drawn into the interface from the XML file. In other words, the website is not constructed from a static set of files, with the transcript and metadata for each interview having been

reformatted and included as part of the coding of its own separate webpage. Rather, it is generated dynamically: there is a single interview interface page that is automatically populated with content from the DECTE XML file associated with the particular interview requested by the user. The corresponding sound file is also automatically made available in the audio player, since it is named in the metadata of the XML file. As indicated in Sect. 2, another crucial way in which the functionality of the interview interface page relies on the format and content of the XML files is the time alignment. By virtue of the time anchor tags that appear in the XML files, the interview transcripts and audio that are loaded into the interface are linked. This means that clicking on any section of text in the interview transcript panel will start playback at the corresponding point in the audio, or jump forwards or backwards to that point if the recording is already playing. In Fig. 7.4, the section of text that begins at the 1-minute mark, with ‘between the Mackems’ has been selected. Consequently, the audio has begun to play from that point and the text of the 20-second segment from that time anchor to the next is highlighted in pale grey.

In Sect. 2 we referred to the ethical considerations that needed to be addressed in preparing the corpus transcriptions and audio files, such as anonymizing them so that no private individuals could be identified. Fig. 7.5 illustrates how this was done: the names of the teachers that the participants in this interview refer to have been replaced with ‘(NAME)’ in the text.

This substitution is not generated dynamically in the interview interface, but is a feature of the underlying DECTE XML files, since these can of course also be downloaded and used separately. For the same reason, the names are blanked out in the associated audio as part of the preparation of the DECTE sound files. Another ethical issue—that is, the fact that parts of some interviews may be considered inappropriate for a general audience—was handled in a different way. In this case, the aim was to block potentially sensitive material, such as words that might be classified as obscenities, from appearing in the open access context of the *Toon* website interface, without redacting the XML transcriptions or removing anything from the sound files, in order to avoid the need to create any separate version of the files that contained different iterations of the interview material. The answer was to add TEI XML tags to the

The Talk of the Toon

Home About Interview Index Themes Quizzes Schools Top Stories Introduction to North East Dialects Links

Archive Interview: Y10i007

Return to: [Theme Results](#) | [Interview Index](#)

For a guide to the layout of this interview page and how to use it, [click here](#).

decten2y10i007audio

0:16:20 0:30:00

Interview Transcript

Speaker 2: Erm *(pause)* *(cough)*

Speaker 1: A teacher? Or?

Speaker 2: There's only *(pause)* two really good History teachers in our *(pause)* erm school

Speaker 3: (NAME) and

Speaker 2: (NAME).

Speaker 3: Ah Yeah.

Speaker 2: Yeah I'm not, don't worry, (NAME)'s a pile of [REDACTED].

Speaker 3: *(laughter)*

Speaker 1: *(laughter)*

Speaker 2: Er

Speaker 3: Saw him at Gateshead the other night

Speaker 2: Did you?

Speaker 3: The only teacher about

Speaker 2: Yeah, Rick Astley

Speaker 3: Yeah

Speaker 2: Yeah

Speaker 1: Rick Astley *(laughter)*

Speaker 1: interviewerY10i007

Speaker 2: informantY10i007a

Age Group: 16-20

Gender: Male

Residence: Tyneside - Newcastle

Education: Higher Education

Occupation: University Student

Speaker 3: informantY10i007b

Age Group: 16-20

Gender: Male

Residence: Tyneside - Newcastle

Education: Higher Education

Occupation: University Student

Themes

Click a theme in the menu below to highlight related keywords in the transcript.

- Education
- Family & Relationships
- Entertainment & Culture
- Religion & Festivals
- Transport
- Community, Politics, Law & Order
- Home
- Clothes & Fashion
- Sport
- Nightlife
- Media & Communication

Fig. 7.5 The interview interface page for decten2y10i007, showing the anonymization of personal names

sensitive material that could then trigger a blocking mechanism written into the code of the *Toon* website's interview interface. An example of this is seen in Fig. 7.5, where a grey box appears in one of the utterances by Speaker 2 (informantY10i007a). The XML version of this section of the transcription is shown in extract (3). This reveals that Speaker

2 in fact says '(NAME)'s a pile of shite'. The presence of the <index indexName="obscenity"> tag that marks the word *shite* is picked up as part of the dynamic process of loading the XML file of the interview into the interface, and consequently the (potentially) offending word is obscured. Working in conjunction with the time anchor tags, the occurrence of the obscenity tag also blocks the corresponding 20-second section of the audio from playing.

(3) An extract from `decten2y10i007.xml`

```
<u who="#informantY10i007b"> (NAME) and </u>
<u who="#informantY10i007a"> (NAME). </u>
<u who="#informantY10i007b"> Ah Yeah. </u>
<u who="#informantY10i007a"> Yeah I'm not, don't worry,
(NAME)'s a pile of <index indexName="obscenity"><term>
shite.</term></index> </u>
<u who="#informantY10i007b"> <vocal><desc>laughter</desc>
</vocal> </u>
```

This obscenity issue came to light when we were discussing the design of *The Talk of the Toon* website with members of the DECTE Advisory Board and with English language teachers from a number of local schools who were attending a continuing professional development (CPD) event that we organized shortly after the DECTE project began.¹⁸ This event consisted of workshops that dealt with topics covered in General Certificate of Secondary Education (GCSE) and A-level English Language courses, including aspects of grammar, language acquisition, language variation and change, and discourse analysis, in a manner not dissimilar to the workshops described in Cheshire and Fox (this volume).¹⁹ The sessions focused on sharing information about recent scholarship on these topics,

¹⁸ The Board included academics from other universities, schoolteachers, museum staff, a senior examiner of Advanced level (A-level) English Language and a member of a local heritage group. For full details of the membership of the Advisory Board, see the DECTE website (<http://research.ncl.ac.uk/decte/people.htm>).

¹⁹ GCSE examinations are taken in England, Northern Ireland and Wales post-16 and students usually sit A-levels during the subsequent two years prior to entering university, for which they are a mandatory qualification.

and on providing the teachers with materials that they could then use in planning lessons and for classroom activities, such as worksheets, reading lists of key articles and information on relevant web resources. From the DECTE project's perspective, the event was a chance for us to get a sense of how teachers might be able to use a resource like *The Talk of the Toon* and to discuss in some detail the initial plans we had for the design, organization and content of the website. The obscenity issue arose in the context of this discussion, and specifically in relation to two questions: (1) how to deal with the fact that school website access is sometimes controlled by software that prohibits the download of unsuitable material and (2) how best to ensure that the material we created was as flexible as possible regarding the target audience so that it was suitable for children of all ages and could be used at all levels of education, from primary to A-level. The teachers were strongly of the view that we should not simply leave out an interview altogether because it had some material which was unsuitable for younger children. Adopting the obscenity XML tag described above enabled us to address this issue without reducing the number of available interviews. The teachers were also keen that the unexpurgated versions of the interviews should be available to older age groups, especially A-level students who might wish to use them for projects (on swearing, for instance) and this became part of the motivation for producing the previously mentioned plain text versions of the transcriptions, which are available through the DECTE research website.

In conjunction with the project's Advisory Board members, the teachers who participated in this event, together with others who attended further outreach activities organized in the School of English Literature, Language and Linguistics at Newcastle University, such as the annual A-level English Language study day, were also instrumental in helping to develop one of the key features of the *Toon* website, namely, the themes. As the screenshot in Fig. 7.4 illustrates, the 17 themes selected appear on the right-hand side of the interview interface pages, under the speaker panels. Clicking on any one of these themes highlights related keywords in the transcript of the relevant interview. For example, Fig. 7.6 presents again the extract from near the beginning of the Y10i009 interview illustrated in Fig. 7.4, but this time with the Sport theme selected, which results in keywords such as *football* being highlighted in the text.

The screenshot shows the 'The Talk of the Toon' website interface. At the top, there is a navigation bar with links: Home, About, Interview Index, Themes, Quizzes, Schools, Top Stories, Introduction to North East Dialects, and Links. The main header features the site logo and a collage of images related to the North East region.

The main content area is titled 'Archive Interview: Y10i009'. Below the title, there is a link to 'Return to: Theme Results | Interview Index'. A sub-header reads: 'For a guide to the layout of this interview page and how to use it, [click here](#).' Below this is an audio player for 'decten2y10i009audio' with a progress bar showing 0:01:04 / 0:29:57.

The central part of the page is the 'Interview Transcript', which contains the following text:

Speaker 2: Well I work with Mackems

Speaker 4: *(interruption)* *(unclear)* and people from Jarrow

Speaker 3: I -- I think *(interruption)* th- th- they're just as good I -- I mean yo-yo- you're -- you're talking about like de- distinguishing between the Mackems

Speaker 2: Yeah I'm talking about in **football** terms.

Speaker 4: Ah

Speaker 3: In **football** *(interruption)* terms

Speaker 2: *(interruption)* In **football** terms er we're -- they're at each other's throats

Speaker 3: Well just find that

Speaker 2: But in if you were to go over and talk to them normally if they didn't know what you were they'll talk to you

Speaker 1: Yeah

Speaker 2: And we'll -- and they'll talk to us I mean

Speaker 1: So is it -- because I'm aware of it but I didn't -- is it a massive rivalry or?

Speaker 3: No

Speaker 1: No

The right sidebar contains speaker information for five speakers:

- Speaker 1: interviewerY10i009
- Speaker 2: informantY10i009a
Age Group: 41-50
Gender: Male
Residence: Tyneside - Newcastle (born in North Shields, North Tyneside)
Education: Left school at 16
Occupation: Driver
- Speaker 3: informantY10i009b
Age Group: 51-60
Gender: Male
Residence: North Tyneside - North Shields
Education: Left school at 16
Occupation: Joiner
- Speaker 4: informantY10i009c
- Speaker 5: informantY10i009d

Below the speaker information is a 'Themes' section with a menu of themes. The 'Sport' theme is selected and highlighted in a dark grey box. Other themes include Language & Identity, Community, Politics, Law & Order, Work & Industry, and Nightlife.

Fig. 7.6 The interview interface page for decten2y10i009, with the Sport theme selected

This highlighting of thematic keywords relies on the same sort of tagging that is used to identify potentially offensive words, as described above: sports-related words such as *football*, for example, are marked with the <index indexName="sport"> tag in the underlying XML files. The decision about which themes would be included and which keywords would

The image shows a screenshot of a website interface. On the left is the 'Themes: Entertainment & Culture' page. It features a vertical navigation menu with categories like 'Family & Relationships', 'Health & Aging', 'Community, Politics, Law & Order', 'Religion & Festivals', 'Language & Identity', 'Education', 'Work & Industry', 'Money & Shopping', 'Media & Communication', 'Transport', 'Entertainment & Culture', 'Sport', 'Holidays', 'Nightlife', 'Food & Drink', and 'Clothes & Fashion'. The main content area shows 'Pictures: 72' and 'Videos: 11'. Below this is an 'Interviews: 67' section with a list of four interviews:

Rank	Interview ID	Year	Speaker(s)	Age Group
1	Y88001	(2000s)	speaker(s): female	41-50 years old
2	TL5506	(1960s)	speaker(s): female	71-80 years old
3	Y87013	(2000s)	speaker(s): male	61-70 years old
4	PVC07	(1990s)	speaker(s): female	51-60 years old

Below the list is a 'Filter Interview Results' section with checkboxes for 'Interview Date', 'Gender', and 'Age'. The 'Age' section includes checkboxes for 14-20, 21-30, 31-40, 41-50, 51-60, and 61-70. On the right side of the screenshot are two sub-pages. The top one is 'Picture Wall: Entertainment & Culture', which displays a grid of 12 small images. The bottom one is 'Picture Archive: Entertainment & Culture', which shows a larger image of a classical building facade and a text box describing the building's history: 'Newcastle's Theatre Royal opened in 1837. The theatre holds an annual residency by the Royal Shakespeare Company and is also the regional home of the National Theatre and Opera North.'

Fig. 7.7 The main Themes page, showing the results for Entertainment & Culture (*left*) and the Picture results for the Entertainment & Culture theme (*right*)

be selected to represent each theme was taken in consultation with the teachers and our Advisory Board members. The intention was to provide users with a pathway that helps them interactively explore the ways in which people in the North East of England talk about their opinions and experiences on a wide range of matters, leading users to the narratives that are most closely related to the topics that interest them.

The highlighting of keywords in a selected interview is not the only opportunity users have to engage with the 17 themes. They are also at the heart of the multimedia content that the *Toon* archive contains. The left-hand side of Fig. 7.7 shows the website's main Themes page. By selecting one of the themes, the user is shown a summary of the relevant pictures, video clips and interviews from the website's archive. For example, in Fig. 7.7, the Entertainment & Culture theme has been selected and the results panel reveals that there are 72 pictures and 11 video clips related to this theme, and that 87 of the corpus interviews contain relevant keywords. Clicking on any of the links in the results panel allows the user to explore this content further. Selecting the image results takes the user to the 'picture wall' of the theme in question, showing all of the images associated with it, each of which can then be selected for further examination. The right-hand side of Fig. 7.7 shows the Entertainment &

Culture picture wall, and below that an individual result, with a larger version of the image, a brief description and details of the source.

In total there are currently 526 images in the archive, some associated with more than one of the 17 themes. Around half of these, mostly depicting modern scenes, are freely available Creative Commons licensed images that were found posted on photo-sharing sites such as Flickr.²⁰ The other half are drawn from the archives of one of DECTE's project partners, Beamish, The Living Museum of the North.²¹ For the most part, these are historical images dating back as far as the late nineteenth century, thus giving us photographic material that illustrates life in the North East across the full span of time from the birth in 1895 of the oldest interviewee in the TLS subcorpus of the 1970s, to the present day and the region as it is known to the participants recorded in the NECTE2 interviews. As a parallel to each theme's picture wall, there is a video wall page containing thumbnail links to the selection of video clips that are connected to that theme.²² There are currently 55 clips in total in this part of the archive, all consisting of short reports from Tyne Tees news programmes of the 1960s and coming from the collections of the North East Film Archive (nefa).²³

The remaining pages of the *Toon* website contain a variety of supplementary materials. The Quizzes page, for example, has links to a set of interactive multiple-choice quizzes with questions mostly about the traditional dialect vocabulary of the region (see Fig. 7.8). These arose from some of the materials that we created for a series of workshops with primary schoolchildren and teachers, held in conjunction with another of our project partners, Newcastle's Discovery Museum.²⁴ The workshops were designed to support teaching and learning in a number of areas, focusing for instance on issues relating to local dialect, history and cul-

²⁰ For more information on Creative Commons licences, see the organization's website (<http://creativecommons.org>). For helping us to find the images and prepare other content for *The Talk of the Toon* website we are indebted to the research assistants on the project: Joanne Bartlett, Claire Childs, Alex Docherty, Kaye Dodds, Jean Price, Nick Roberts and Peter Wilson.

²¹ See <http://www.beamish.org.uk>.

²² See, for example, the video wall for the Entertainment & Culture theme at: http://research.ncl.ac.uk/decte/toon/videos.html?theme=entertainment_culture

²³ See <http://www.northeastfilmarchive.com>.

²⁴ See <https://discoverymuseum.org.uk>.

The screenshot shows two overlapping digital interfaces. The left interface is titled 'The Talk of the Toon Quiz One' and features a question: 'Which of the terms below is NOT a word for marbles?'. The options are: a. radgies, b. liggies, c. allies, d. moggies, e. glassies. A 'Correct!' message indicates that 'a radgy is an angry or aggressive person' and lists other words for marbles: 'pappers, penkers (large marbles made of iron or stone) and scudders (the shooting marble)'. The right interface is titled 'The Talk of the Toon' and contains sections for 'Words and Objects' with questions about dialect words, a 'Dialect Words' section with a list of words and their meanings, and another 'Words and Objects' section with questions about conversation topics. It also includes a 'Word Use' table and a 'Answers' section.

Fig. 7.8 A question from one of the *Talk of the Toon* interactive quizzes (left) and materials from the Discovery Museum workshops (right)

tural heritage, the characteristics of spoken versus written language and the way in which language use changes in relation to different contexts and audiences. To address these topics we drew on a combination of material from the corpus and objects from the Museum's collections to create a range of activities, including word games, role-playing exercises, and a dialect poetry competition (see Fig. 7.8).

As well as feeding into *The Talk of the Toon's* quizzes, additional activities from these workshops are available for download in the Schools section of the website, together with the materials from other project events for students and teachers. These include the CPD sessions for English language teachers described above. We also ran a workshop for A-level students on using the plain text versions of the corpus transcriptions with the previously mentioned *AntConc* software. There were also sessions for A-level teachers that focused on DECTE during the *Analysing Spoken English* workshops already noted and discussed elsewhere in this volume, by Cheshire and Fox. In addition to the materials that are principally designed for users in the educational sector, there are other parts of the website that are specifically intended to appeal more broadly, namely a short introduction to North East dialects, with information also on the origin of terms for local people, such as *Geordie* and *Mackem*, and a Top Stories page that highlights some of the best narratives from the corpus interviews. The material from these sections also features in a *Talk of the*



Fig. 7.9 *The Talk of the Toon* booklet

Toon booklet and CD package. An extract from the booklet, presenting the same Y10i009 interview that has been seen in its online form, TEI XML file format and plain text version above, appears in Fig. 7.9. As well as using it for promotional purposes, our aim in producing the booklet has been to find a way of reaching those members of the community who may not typically engage with online resources, in part simply as a way of taking the material to them, but also as a way of encouraging them to come to the *Toon* website, to see the greater range of interviews and other resources available there.

4 Conclusion

In this chapter, we have discussed the strategies that were employed in creating the *Diachronic Electronic Corpus of Tyneside English*. The aims of the project were to ensure the longer-term preservation and sustainability of the resource, and to develop an associated public engagement programme that would make the corpus accessible and relevant to a diverse range of non-academic user groups in the wider community. We have argued that a crucial factor in successfully achieving these aims has been the adoption of an XML file format as the basis for the underlying corpus architecture, and

in particular one that conforms to the widely used standards established by the Text Encoding Initiative for the digital representation of documents. The TEI XML format has a number of advantages in terms of the basic management of the files, including the ease with which it can be maintained and updated, and the fact that it affords a high degree of compatibility, thus facilitating integration with other data sets, software packages and electronic resources. However, it has also proven to be an essential element in the effective dissemination of the material outside the sphere of academic research and the establishment of DECTE as a public corpus that appeals to different kinds of users and has applications in a broad range of contexts (see Kretzschmar et al. 2006: 179–82). The project's open access website, *The Talk of the Toon*, exploits the TEI XML file format to present the corpus interviews through an interface which encourages users to interact with them as an archive of local oral history that reflects the North East's linguistic and cultural heritage. To add depth to the archive, the *Toon* website also has a multimedia dimension, linking pictures and video clips to thematically related discussions and narratives of personal experience in the corpus interviews. The development of this and other aspects of the site benefitted enormously from the fact that prospective users were involved from the outset in key decisions about the overall look and organization of the material, effectively participating in a process of co-design. Working with potential users in this way helped us to fulfil our aim of fostering mutually beneficial and productive relationships with a wide range of users across the community, in line with the concept of linguistic gratuity advocated by Wolfram (1993) and the concerns that currently inform the public engagement and impact agendas in UK higher education.

Resources

Books and Articles

Allen, Will, Joan C. Beal, Karen P. Corrigan, Warren Maguire and Hermann L. Moisl. 2007. A linguistic 'time-capsule': The Newcastle Electronic Corpus of Tyneside English. In *Creating and Digitizing Language Corpora, Volume 2: Diachronic Databases*, eds. Joan C. Beal, Karen P. Corrigan and Hermann L. Moisl, 16–48. Basingstoke: Palgrave Macmillan.

- Beal, Joan C., Karen P. Corrigan, Adam J. Mearns, and Hermann L. Moisl. 2014. The *Diachronic Electronic Corpus of Tyneside English*: annotation and dissemination practices. In *The Oxford Handbook of Corpus Phonology*, eds. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 517–533. Oxford: Oxford University Press.
- Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, Ben Rampton and Kay Richardson. 1997. Ethics, advocacy and empowerment in researching language. In *Sociolinguistics*, eds. Nikolas Coupland and Adam Jaworski, 145–162. Houndmills: Macmillan. (Originally published in *Language and Communication* 13(2): 81–94 in 1993.)
- Jones-Sargent, Val. 1983. *Tyne Bytes: A Computerised Sociolinguistic Study of Tyneside*. Frankfurt am Main: Peter Lang.
- Kretzschmar, William A., Jean Anderson, Joan C. Beal, Karen P. Corrigan, Lisa-Lena Opas-Hänninen and Bartek Plichta. 2006. Collaboration on corpora for regional and social analysis. *Journal of English Linguistics* 34(3): 172–205.
- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- . 2004. Ordinary events. In *Sociolinguistic Variation: Critical Reflections*, ed. Carmen Fought, 31–43. Oxford: Oxford University Press.
- . 2006. Narrative preconstruction. *Narrative Inquiry* 16: 37–45.
- Milroy, James, Lesley Milroy and Gerry Docherty. 1997. Phonological Variation and Change in Contemporary Spoken British English. ESRC, Unpublished Final Report, Dept. of Speech, University of Newcastle-Upon-Tyne.
- Milroy, Lesley. 1987. *Observing and Analysing Natural Language: A Critical Account of Sociolinguistic Method*. Oxford: Blackwell.
- Milroy, Lesley, James Milroy, Gerry Docherty, Paul Foulkes, and David Walshaw. 1999. Phonological variation and change in contemporary English: evidence from Newcastle-upon-Tyne and Derby. *Cuadernos de Filología Inglesa* 8(1): 35–46.
- Pellowe, John, and Val Jones. 1978. On intonational variety in Tyneside speech. In *Sociolinguistic Patterns in British English*, ed. Peter Trudgill, 101–121. London: Arnold.
- Pellowe, John, Graham Nixon, Barbara Strang, and Vincent McNeany. 1972. A dynamic modelling of linguistic variation: the urban (Tyneside) linguistic survey. *Lingua* 30(1): 1–30.
- Strang, Barbara. 1968. The Tyneside Linguistic Survey. *Zeitschrift für Mundartforschung*, NF 4 (Verhandlungen des Zweiten Internationalen Dialektologenkongresses). Wiesbaden: Franz Steiner Verlag, 788–794.

- Watt, Dominic. 2002. 'I don't speak with a Geordie accent, I speak, like, the Northern accent': Contact-induced dialect levelling in the Tyneside vowel system. *Journal of Sociolinguistics* 6(1): 44–63.
- Watt, Dominic, and Lesley Milroy. 1999. Patterns of variation and change in three Newcastle vowels: Is this 'dialect levelling?'. In *Urban Voices: Accent Studies in the British Isles*, eds. Paul Foulkes, and Gerard J. Docherty, 25–46. London: Arnold.
- Wolfram, Walt. 1993. Ethical considerations in language awareness programs. *Issues in Applied Linguistics* 4: 225–255.
- Wolfram, Walt, Jeffrey Reaser, and Charlotte Vaughan. 2008. Operationalizing linguistic gratuity: from principle to practice. *Language and Linguistics Compass* 2(6): 1109–1134.

Websites, Software and Online Resources

- AntConc*: Anthony, Laurence. 2015. *AntConc* (Version 3.4.4) Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net/software/antconc> (accessed 7 August 2015).
- DECTE/*The Talk of the Toon*: Corrigan, Karen P., Isabelle Buchstaller, Adam Mearns and Hermann Moisl. 2012. *The Diachronic Electronic Corpus of Tyneside English* (DECTE) and *The Talk of the Toon*. Newcastle University. <http://research.ncl.ac.uk/decte> and <http://research.ncl.ac.uk/decte/toon> (accessed 7 August 2015).
- ELAN: Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, <https://tla.mpi.nl/tools/tla-tools/elan> (accessed 7 August 2015).
- ENROLLER: Enhanced Repository for Language and Literature Researchers, <https://enroller.nesc.gla.ac.uk> (accessed 7 August 2015).
- NCCPE: The National Co-ordinating Centre for Public Engagement, <http://www.publicengagement.ac.uk> (accessed 7 August 2015).
- NECTE: Corrigan, Karen P., Hermann Moisl and Joan Beal. 2005. *The Newcastle Electronic Corpus of Tyneside English* (NECTE). Newcastle University. <http://research.ncl.ac.uk/necte> (accessed 7 August 2015).
- NECTE2: <http://research.ncl.ac.uk/necte2> (accessed 7 August 2015).
- OTA: *The Oxford Text Archive*, <http://ota.ox.ac.uk> (accessed 7 August 2015).
- RCUK: Research Councils UK, <http://www.rcuk.ac.uk> (accessed 7 August 2015).
- TEI: *The Text Encoding Initiative*, <http://www.tei-c.org> (accessed 7 August 2015).
- TXM: *Textométrie*, <http://textometrie.ens-lyon.fr> (accessed 7 August 2015).

Wordsmith: Scott, Mike. 2012, *WordSmith Tools version 6*, Stroud: Lexical Analysis Software. <http://www.lexically.net/wordsmith> (accessed 7 August 2015).

Xaira: XML Aware Indexing and Retrieval Architecture, <http://xaira.sourceforge.net> (accessed 7 August 2015).

8

Language Learning at Your Fingertips: Deploying Corpora in Mobile Teaching Apps

Seth Mehl, Sean Wallis, and Bas Aarts

1 Introduction

Since 2009, the *Survey of English Usage* (SEU) at University College London has developed a series of knowledge transfer and pedagogical innovation projects for teaching English linguistics to a broad audience. The results of these projects include three language-learning mobile apps: the *interactive Grammar of English* (iGE; Aarts and Wallis 2011), *Academic Writing in English* (AWE; Mehl et al. 2013), and *English Spelling and Punctuation* (ESP; Wallis et al. 2014).¹ In this chapter, we begin by briefly summarizing the history of language teaching, representing the context into which these language teaching apps were introduced. In Sect. 3, we

AWE and ESP were funded by a UCL Teaching Innovation Grant, and iGE was developed as a spin-off from the AHRC-funded projects *Creating a Web-Based Platform for English Language Teaching and Learning* (AH/H015787/1) and *Extending the Englicious Platform for Primary English* (AH/L004550/1). We gratefully acknowledge this support.

¹ The three apps can be accessed online via the *Survey of English Usage*: <http://www.ucl.ac.uk/english-usage/apps>. A fourth app, Grammar Practice for KS2 (GP-KS2), is also available.

S. Mehl (✉) • S. Wallis • B. Aarts
University College London, London, UK

© The Editor(s) (if applicable) and The Author(s) 2016
K.P. Corrigan, A. Mearns (eds.), *Creating and Digitizing
Language Corpora*, DOI 10.1057/978-1-137-38645-8_8

discuss the challenge of delivering educational materials on a smartphone platform. In Sects. 4–6 we then discuss the three apps in detail. In Sect. 7 we explore the broader implications of these apps with a particular focus on the pedagogical issues and technical challenges that the apps raise.

2 A Brief History of Language Pedagogy

Around 50 years ago, foreign language skills in schools were often taught using so-called ‘drills’. These were patterns of linguistic structures that students were asked to repeat with the aim of consolidating them into their communicative repertoire. McCaul (1973: 6) defended teaching through drills, writing:

Students enjoy drills and the teacher should make the drills contextualized, situational and interesting. The primary aim in the use of drills is for the student to be able to transfer his drill habits into his conversation. In order to accomplish this it is necessary for the student to be intellectually and, if possible, emotionally involved in the activity.

While McCaul’s aim of contextualized education seems well rounded, it is not so evident that students enjoyed drills. One of the authors of the present chapter remembers learning French at school through so-called *exercices structuraux*, and this not being a particularly pleasurable experience. The grammatical patterns to be learnt were presented embedded in blocks of ten stale and unexciting sentences which had to be learnt by heart. This particular type of exercise is described by McCaul (1973: 6) as follows:

The Straight Pattern Practice. The teacher drills the grammatical pattern she wishes to emphasize until it is learned. She then makes a vocabulary change while leaving the grammatical structure the same.

Even if this process worked, it is difficult to see how students can be encouraged to be intellectually and emotionally involved in this kind of behaviourist learning. Teachers soon realized that drills were not only uninspiring, they also did not teach students to use, reflect on, or

analyse language in natural linguistic settings.² Moreover, as in other fields of linguistics, language pedagogy in the 1970s reacted strongly to Noam Chomsky's research: generative grammar and its decontextualized approach inspired educationalists to focus anew on contextualized communication (see Prodromou 1996; Gilmore 2007). The drill-based, behaviourist phase of language teaching gave way to Communicative Language Teaching (CLT) and Whole Language Teaching, and these movements rallied around the use of authentic language in the classroom (see Widdowson 1998; Larsen-Freeman 2000; Gunderson 2009).

The rise of the communicative approach can ... be seen as the response of the language teaching profession to their new situation, and a recognition of the inadequacy of traditional 'grammar/translation' methods, and also of the 'structural' methods of the 1950s (which stressed speaking and listening but relied heavily on meaningless pattern drills and repetition), to meet the needs of their new publics. Fortunately, at the time when there was a will for change, a range of new ideas in different branches of linguistics began to offer a range of possible new solutions. (Mitchell 1994: 34)

Contemporaneously, Widdowson (1979: 75) described English textbooks designed for student use as 'contrived' and decontextualized from actual communication. Communicative methods were in turn criticized (Swan 1985; Bax 2003) and a new debate ensued (see Widdowson 1985). Like Widdowson (1979), Willis (1990: 127) observed a 'contrived simplification of language in the preparation of materials', a stance echoed by Goodman and Freeman (1993). However, the notion of 'contrived language' in classroom drills as opposed to 'authentic language' in context remained vague, and these informal observations led to no rigorous studies on the precise linguistic differences between putative 'stilted examples' and 'contextualized language', much less to research on the pedagogical efficacy of either. More recent pedagogical research, in contrast to the former debate between 'authentic language' and 'stilted

² The British comic group Monty Python made fun of the drill-based approach in a sketch in which a foreigner walks into a shop holding a Hungarian phrase book, and says to the shopkeeper: 'I will not buy this record; it is scratched'. The shopkeeper replies: 'No, no, no, this is a tobacconist's'. The visitor, experiencing an *aha-Erlebnis*, then says: 'Ahh, I will not buy this tobacconist's; it is scratched'.

examples' as key to language learning, often highlights the diversity of factors that impact on successful learning, including the surrounding environment and student motivation. Such new research generally avoids claims that a particular type of language example, type of exercise, textbook or other resource can be uniquely effective independent of those diverse factors (see Royer et al. 1984; Gass 2003; Gass and Torres 2005; Grabe 2009). Thus, ill-defined claims regarding contrived drills and contextualized language have begun to give way to more rigorous psychosocial and pedagogical studies on the web of factors that facilitate student achievement. Research on corpora in pedagogy has opened up a new and different perspective on the problems that may arise from using language from natural contexts in classroom resources; we return to the question of corpora in pedagogy in Sect. 7.

The teaching of grammar to native speakers virtually disappeared in the UK in the first half of the twentieth century, mainly because the teaching of literature and writing *sans* grammar was dominant (Hudson and Walmsley 2005). After the second world war, a number of government reports recommended the teaching of grammar in schools, but the cultural rise of CLT and Whole Language Teaching often went hand in hand with an objection to the explicit teaching of grammar, on the rather shaky grounds that teaching 'grammatical signification' as such was inherently decontextualized from practical communication (Widdowson 1979: 75). UK policies developed, and a National Curriculum for the teaching of English was introduced in 1988, followed by a National Literacy Strategy in 1999. A new National Curriculum for 2014 requires a relatively complete understanding of English grammar by Year 6 in British primary schools.

The SEU has been at the vanguard of grammatical research in English for decades, and has actively deployed its research expertise in undergraduate and postgraduate English grammar courses. The Survey was the first European research group in corpus linguistics and began the process of compiling the first corpus of spoken English ('The Survey Corpus') in the 1950s. This archive made possible the creation of the corpus-based grammar of English, *A Comprehensive Grammar of the English Language* (Quirk et al. 1985). The Survey's research since that time has continued to centre on corpora and language in use. That focus, central to corpus lin-

guistics, has informed our teaching, and is particularly relevant at a time when public debate in English language education has become dominated by a government insisting on a renewed emphasis on explicit and prescriptive rules surrounding spelling, punctuation and grammar.

Since the publication of our first parsed corpus, the *British component of the International Corpus of English* (ICE-GB; Nelson et al. 2002) in 1998, a revolution in electronic access in education has taken place worldwide. The SEU's *Internet Grammar of English* (IGE),³ published in 1996, represented the first wave of this revolution. We are currently seeing a second wave, characterized by portable electronic textbooks and mobile applications ('apps'). This second wave appears to have taken place with considerable public debate, but remarkably little academic analysis.

The rapid uptake of technology has not been uncontroversial. The initial reaction of many schools to mobile phones in the classroom was to ban them as a distraction, rather than view them as a potential adjunct to classroom teaching. The rise of smartphone apps raises questions of educational authority as well as classroom control. Mobile phones are typically private devices for students, as distinct from terminals in libraries or schoolrooms, where access to particular resources can be managed by the school.

This section has described the context of our language app programme at the SEU. The next section includes a brief overview of the apps, followed by a discussion of each app in turn (Sects. 4–6). Sect. 7 draws out three themes from these case studies: the role of interactivity (Sect. 7.1), the application of natural language and the role of grammar (Sect. 7.2), and the problem of selecting corpus examples for the purposes of linguistic pedagogy (Sect. 7.3).

3 Mobile Apps for Language Learning

Our three case studies deploy corpus data in mobile apps. The apps are self-contained resources that cover distinct topics:

³ See <http://www.ucl.ac.uk/internet-grammar>. IGE was funded by JISC JTAP-49.

- (a) The *interactive Grammar of English* (henceforth iGE; see Aarts et al. 2012) is comparable to a full undergraduate course in English grammar. Alongside a step-by-step structured introduction to grammatical terminology and analysis, the app also includes an extensive searchable glossary that allows students to look up unfamiliar terms, as well as a set of nearly 40 exercises drawing on a bank of hundreds of corpus examples that allows students to practise what they have learned.
- (b) *Academic Writing in English* (AWE) is a tool for learning the principles of writing essays and dissertations for undergraduate and post-graduate contexts. It covers academic writing from the abstraction of categories of critical thinking and argument construction, to awareness of academic register and appropriate word choice, and lower-level issues such as presentation and referencing conventions.
- (c) *English Spelling and Punctuation* (ESP) contains two short courses. The spelling course begins with basic spelling rules and proceeds to a series of interactive exercises on frequently misspelled words and words that represent exceptions to the standard rules, ordered by decreasing frequency and increasing specificity. The punctuation course helps students learn standard punctuation in goal-driven contexts, and explains the logical foundations of the choices made along the way. The punctuation app also includes a thorough reference guide covering the major punctuation marks.

The three apps can be seen as occupying a spectrum from the most interactive and least book-like ESP to the least interactive and more book-like AWE, with iGE falling in the middle. This spectrum is discussed further in Sect. 7.

The three apps could have been implemented as websites, but in website form they would not have been as readily portable. A prototype, the *Internet Grammar of English* (IGE), pre-existed the app by over a decade (see Sect. 4). However, converting a web or web-style resource to an app format presents a number of challenges, both technical and pedagogical.

- (a) *Independence*. An important by-product of mobility is intermittent (or expensive) Internet access. Whereas a web user has access to search engines for any difficult term, app designers need to ensure that the

- entire app, including supporting material such as glossaries, is sufficient for learners' needs without requiring Internet access, so that the app itself is useful even when the user is not online.
- (b) *Screen size.* As software, apps must be designed for a full range of hardware, and app designers must be sensitive to the specifications of that hardware. Apps for small-screen devices, such as smartphones, need simple explanations, short sentences and paragraphs fitting on a single screen (for example, the Sony Xperia Mini is a device with a 2.55 inch screen of 240 × 320 dots; see Fig. 8.1). Increased resolution on 'high end' devices means smoother text and images, but not necessarily larger screen sizes. Both explanatory text and corpus examples must be viewable without scrolling, and the user must have a sense of coherence across screens or they will rapidly lose their place.
- (c) *Multiple form factors.* App designers have to consider a number of different variables, including different device sizes, from small screen phones to tablets (Fig. 8.1), some have physical keyboards, and some 'clamshell' devices even have dual screens. Even smaller devices like the Galaxy Gear 'smart watch' present a further challenge if they are to be targeted.
- (d) *Touch.* Touchscreens have decluttered phone interfaces and dispensed with cursor keys and trackballs, but they present their own



Fig. 8.1 The form factor challenge: iGE on a tiny Sony Xperia Mini Android phone (left) and Motorola Zoom tablet (right)

interface design challenges. Whereas mouse cursors are fine and precise, touch areas and gestures are necessarily imprecise, and any interface that fails to plan for this will not succeed. We return to this issue in Sect. 7.

- (e) *Additional smartphone features.* App development in general offers additional options such as engaging with global positioning facilities, tilt sensors, and vibration. The three apps discussed here use vibration in the interactive exercises, but no additional options.

4 The *Interactive Grammar of English* (iGE)

4.1 Introduction to iGE

iGE is an app consisting of a complete course in English grammar, supplemented by a searchable glossary and 36 interactive exercises. In this section we present a thematic overview of the app, which is described in detail in Aarts et al. (2012).

The exercises in iGE draw on a bank of over 600 sentence examples drawn from our ICE-GB corpus and the course content was developed from the *Internet Grammar of English* (IGE). IGE was the first complete grammar on the Internet aimed at moderately advanced learners of English who could benefit from a succinct, but reliable, account of English grammar with exercise material. (Similarly, iGE is the first reference grammar course on the iPhone App Store and the Android Google Play store.)

The course covers *word classes* (nouns, determiners, verbs, and so on); *phrases, clauses and sentences*; and *grammatical functions* (subjects, objects, and so on). A number of other grammatical concepts, such as *semantic roles* and *extraposition*, are introduced as required. It is structured in a linear fashion, so that the student is encouraged to read it from start to finish (see Fig. 8.2).

In developing the iGE app from the IGE website, the course was first thoroughly rewritten to bring it up to date with contemporary grammatical thinking. The content itself was also redesigned to make it easy to read

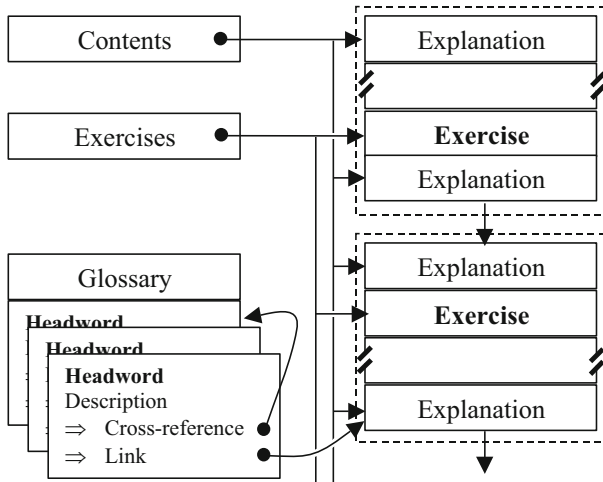


Fig. 8.2 Sketch of app structure: navigation tools (*left*) and course content (*right*). The course consists of a series of ‘chapters’ containing sequential explanation and exercise pages

on small-screen devices (see Sect. 3 above). Explanations and examples were shortened for readability, and the user interface was completely overhauled for touchscreens. The screen display was initially optimized for ‘phone’ and ‘tablet’ configurations, and this customization was further extended for the wide range of particular Android devices.

4.2 Navigation in iGE

In order to make navigation as simple as possible, iGE was given three navigation menus: Contents, Exercises and Glossary. Contents lists the content of the app, including material such as About and Further Reading pages that are not really part of the course material, and links to every page in the course itself; Exercises lists the exercises, and displays the user’s score for each exercise, summarizing their progress; and Glossary launches the glossary tool, which also serves as an index. This navigational structure, which is common to all three apps, is shown in Fig. 8.2.

The Glossary cross-references itself and also indexes pages in the course, acting as a powerful navigational tool. For example, the entry for *Noun*

gives a short definition, offers cross-references to other relevant glossary entries, and links straight to the course material on nouns. As one might expect, the reverse is also true, so the user can look up *Noun* in the glossary whenever it is reintroduced throughout the course. This constitutes an extremely practical, complex, yet user-friendly network of connections throughout the app. As an additional mode of navigation, a student can open a keyboard and type a search string.

4.3 Exercises in iGE

In the iGE app, an important change from the website radically alters the way that exercises work ‘under the hood’. In the *Internet Grammar of English* website, exercises were essentially fixed around a small number of selected examples. Commentary and explanation could then be hand-crafted around these examples. However, the high cost of this approach meant that the number of individual examples used per exercise would inevitably be very small. Repeating the exercise meant seeing the same example again. A student could get the answer correct by simple recognition, not by remembering the underlying principle.

Employing corpus databases for learning includes the potential for benefits of realism and scale. Naturally occurring examples are easily understandable to the student and the grammatical principles they embody may be applied by the student in their own writing, or in other contexts of natural use.

Scale is at least as important as realism. The app allows students the opportunity to carry out the same exercise with many different examples. If varied examples are available, the student can begin to recognize the underlying principle being taught and learn to apply what they have learned in a variety of contexts. One of the benefits of classroom concordancing (see Sect. 7.3) was the availability of many examples to reinforce learning. Unlike classroom concordancing, in iGE students are also taught the explicit rules to apply, and examples have been preselected for pedagogical value.

The software developed for iGE includes two types of multiple-choice exercises (single-answer ‘radio button’ and multiple-answer ‘check box’ exercises), and two types of selection exercises (the first type for selecting individual words in the text; and the second type for selecting a range within the text, as, for example, identifying a noun phrase subject). The software selects examples for a particular exercise test from a bank of preselected corpus examples. It employs a random ‘shuffle’ algorithm that allocates examples to each exercise run in a structured way. This has two aims: to avoid the same example coming up twice in succession, and, in the case of radio-button exercises, attempting to ensure an even distribution of answers, so not all questions will have the same answer.

Although handcrafted responses from the app to all 600+ examples are not feasible, and students may answer exercises in a range of incorrect ways which might conceivably require a number of different explanations from a teacher, the app can provide some targeted feedback. The best way of explaining the type of feedback possible is by an example.

The following output responds to an exercise to identify which of the following nouns are count nouns: *age*, *health*, *middle*, *awkwardness* and *music*. The app first outputs the user’s response and scores each in turn, and then summarizes the particular examples as follows:

Count nouns usually have different singular and plural forms. In the singular they usually take *a/an* before them. So two are count nouns: *age* and *middle*.

Non-count nouns may be considered to refer to indivisible wholes. They do not normally have plural forms, or take *a/an*. Thus three refer to non-count nouns: *health*, *awkwardness* and *music*.

This response integrates the test words in both categories (*{age, middle}*, *{health, awkwardness, music}*) into a textual explanation that refers back to the principle being taught.

4.4 Gamification in iGE

A final aspect of the development of iGE and its sister apps that is worth some comment is a move towards ‘gamification’—the tendency to use techniques drawn from the computer games industry in learning tools.

One of the key goals in designing iGE for a mobile phone platform was to make the app more approachable and entertaining than a conventional grammar textbook. In addition to simplifying the text and employing bright colours, we added sound, visual effects and even vibration to the exercises. We added an optional clicking clock (top right in Fig. 8.3) to allow students to compete against time, and to permit students to improve on 100 per cent success by answering questions faster.

As befits a personal device, scores are essentially unique to the student, and data is only stored locally on the mobile device. Although the technology exists for users to share data by wi-fi, phone (3G, 4G) or peer-to-peer radio connection (‘Bluetooth’), these can incur unexpected costs. However, we envisage that future enhancements may include the ability for students to challenge each other remotely, or for teachers to guide students’ reading through the app for whole-class study.

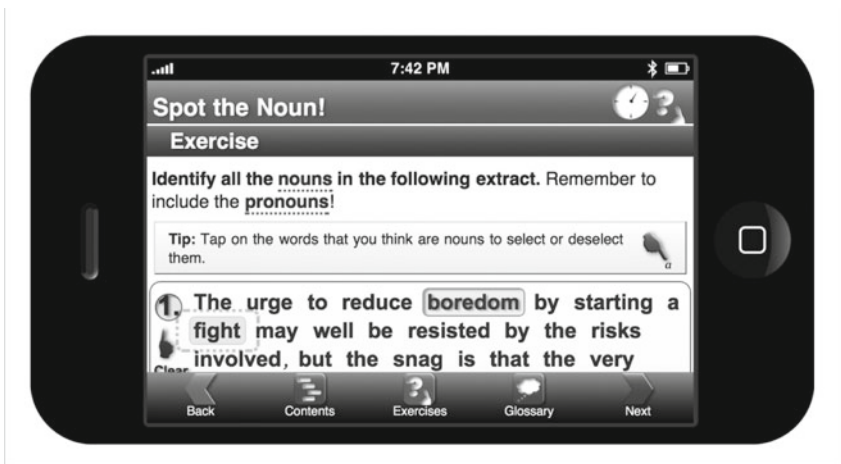


Fig. 8.3 A ‘spot the noun’ exercise in iGE (iPhone). The exercises can be performed in horizontal or vertical orientation

5 *Academic Writing in English (AWE)*

5.1 Introduction to AWE

Following the positive response to iGE, the Survey began a project to develop two additional English language apps aimed at an undergraduate student audience. Thus, *Academic Writing in English (AWE)* and *English Spelling and Punctuation (ESP)* were born. Each of the two apps was built on the design template established by iGE. This meant that we already had a consistent template for presenting explanatory text and examples, as well as software to present the glossary and exercises, navigate the app, and so on, although the subject matter of each app raised new challenges. AWE is discussed here, and ESP is discussed in Sect. 6.

AWE is designed to be a full self-study course that explores what academic writing is and why we do it, building from a foundation of critical thinking and research skills. AWE thus approaches academic writing in a goal-oriented way. According to AWE, when we engage in academic writing we aim to convey our own critical thinking skills, our own thoughts and arguments, as we respond to and engage with the work of others. Our envisaged ideal user is an undergraduate sitting at a computer writing an essay, and using AWE to stimulate their thought processes, or to check their essay for completeness. Although it is linearly structured, we expect students to dip in and out. Ease of navigation is therefore essential.

AWE is organized under the headings below, which take the reader from first principles to a completed piece of coursework.

- (a) *Foundations* explains how to interpret an assignment question and create an ‘essay plan’. It explains that academic writing may have different purposes and standards according to the discipline. The term ‘essay’ is used generically throughout to cover a range of undergraduate assignment types. These are categorized using the ‘genre families’ typology developed for the *British Academic Writing in English* corpus (BAWE; Nesi and Gardner 2012). From the start, the app takes the student through the research process, including issues of integrity in



Fig. 8.4 Left: An example AWE content page (iPhone). Right: An exercise in AWE on the iPhone to select the most appropriate transition. The option *But also ...* has just been selected

referencing and simple ways of thinking and practice that help students avoid plagiarism. This chapter also introduces techniques for taking notes on sources and incorporating source quotes into essays (summarizing and paraphrasing).

- (b) *Argument* discusses what an academic argument and counterargument look like and introduces the concept of ‘critical thinking’, using the taxonomy of Bloom et al. (1956) as a guide. We also evaluate the BAWE genre families against Bloom et al.’s criteria, to show how critical thinking tasks may differ in different academic writing (see Fig. 8.4).

- (c) *Structure* introduces methods for organizing writing, discusses common frameworks for structuring essays from introduction to conclusion, and then looks at scientific experimental reporting structures and scientific essays.
- (d) *Style* is the main part of the app that discusses language style and register. The chapter contains the majority of the interactive exercises, because these are more straightforward to implement. It covers structural decisions within an essay section, such as how to direct the reader through transitions between paragraphs and sentences. And it discusses the practice of punctuating joins and integrating quotations. The *Style* section ends with a discussion of word choice and some grammatical considerations (prepositional phrases, use of phrasal verbs and disambiguating pronouns).
- (e) *Rhetoric* is a short chapter covering a couple of topics to help advanced writers improve their style. These are the employment of effective analogies and the art of parallel construction. (Later releases may expand this chapter.)
- (f) Finally, *Completion* covers simple techniques for finalizing the essay, checking that it has met the objectives set, revising content and presentation, and finalizing conventions.

In summary, AWE is not a mere bag of stylistic tricks for academic writing, such as cohesive devices or highfalutin vocabulary. Rather, it explains to users from the beginning that academic writing is (to quote AWE) ‘no mere technical skill: it is built on a way of thinking’.

5.2 Interactivity in AWE

This sequential structure means that AWE is the least interactive of all three apps. It can be seen as a kind of innovative e-textbook designed specifically for smartphones and tablets, with a few interactive exercises around focused topics such as vocabulary building or recognizing the difference between language of a high or low register (see, for example, Fig. 8.4). This reduced interactivity reflects the pedagogical aims: not every pedagogical aim benefits from multiple-choice question-and-answer tests, or recogni-

tion exercises (such as ‘Spot the Goal’). The process of academic writing requires the student to engage with multiple levels of thought processes and rapidly prodding the screen may be distracting rather than helpful in many cases. These different thought processes include:

- (a) activation of critical thinking skills;
- (b) engagement with other academic sources;
- (c) nuanced construction of arguments;
- (d) discursive writing strategies at the level of the paragraph and beyond;
and
- (e) linguistic techniques at the level of the sentence or the word, including precision and accuracy in word choice and elegant variation.

AWE is therefore, by necessity, very different from iGE. The latter deals with the sentence alone as the fundamental unit of grammar, and individual example sentences can be easily presented, analysed and explained in interactive examples on a small screen. Application of critical thinking skills to an academic article in a journal, for example, is not so easily presented, much less within an interactive exercise on a pocket device.

Like an e-textbook, AWE is centred around a sequence of primarily text-based chapters. Unlike a typical e-textbook, and like iGE, AWE’s chapters are designed to be small enough, and each paragraph brief enough, to be read comfortably on a very small device. Some additional interactivity is included in what we term ‘tasks’, which are not interactive exercises that provide feedback, but activities and projects that students can perform away from the app. For example, one task asks students to compose counterarguments to a set of given arguments drawn from newspaper opinion columns. Possible answers are revealed (if desired) by the user. These tasks encourage students to engage with all of the app’s content, even that content that cannot be practised using simple multiple choice or identification exercises. Unlike exercises, these tasks are not scored.

At launch, AWE included 10 exercises, mostly concentrated in the ‘Style’ chapter, which focuses on language skills (see, for example, Fig. 8.4). Exercises earlier in the app require students to identify key words in real undergraduate essay questions, and to draw parallels between the tasks required by real undergraduate essay questions and the tasks represented

in a critical thinking skills rubric. Later exercises focus on the level of the sentence, and provide examples of natural language from ICE-GB.

ICE-GB contains 80,000 words of academic writing and 20,000 words of student essays and exam scripts. Using this resource, the app presents to users, in bite-sized chunks, excerpts at both a student and experienced academic level. A larger part of the corpus is also available to expose differences in register, for example, between academic and non-academic writing.

Users are encouraged to apply their developing knowledge about the nature of academic writing to these manageable chunks. For example, students are asked to look at examples from the corpus and identify specific linguistic features as being typically academic or non-academic; or to identify examples of clear and unclear language in context.

Other exercises present examples of common academic terms, particularly polysemous terms, in multiple contexts and ask students to identify appropriate synonyms in each case. Still other exercises provide pairs of sentences from academic publications and ask students to link the sentences together fluently. All of these practical exercises follow from, and explicitly build upon, the foundation of critical thinking skills that underlies good academic writing.

The free app includes 10 exercises with between 25 and 50 examples in each exercise. A further 200 examples of natural language from the corpus will soon be available as an in-app purchase for students to continue to practise their skills.

6 *English Spelling and Punctuation (ESP)*

6.1 Introduction to ESP

The third app in our series is ESP, *English Spelling and Punctuation*, aimed at undergraduate level students. This is really two short courses in one, plus a large interactive test component.

- (a) The *Punctuation* course explains a standard method for punctuating text based on the rigid principles of the Oxford University Press style guide. We do not simply teach this style guide, but discuss why these

principles are proposed. For example, many students are unsure as to when it would be appropriate to use a dash, parentheses, semicolon, or comma. The app clarifies that this may be a question of taste or authorial choice, but it is possible to make principled decisions about such a choice. Punctuation principles are discussed in the context of making a particular type of structural decision in writing, such as creating a list, or commenting on the main action of the sentence. As with all our apps, corpus examples populate the exposition and the occasional test exercise.

- (b) The *Regular Spelling* course discusses morphology and focuses on rules for adding regular suffixes to words. This is an area known to be beneficial for undergraduates who frequently need to be able to add a suffix to a novel technical word, or recognize when such a word is already inflected. Although there are no absolutely reliable rules of spelling, even in this area, there are guiding principles that can be taught. The app contains exercises where the user simply is given the base word and suffix and types the correct complete word.
- (c) Finally, *Spelling Practice* consists of interactive exercises that mimic the situation where a student is struggling to spell a word correctly. We discuss this module in more detail below.

6.2 Interactivity in ESP

In ESP, *Spelling Practice* consists of 200 words organized into two 100-word tiers containing four units of 25 words. These units are ordered in decreasing order of exposure frequency (as attested by the very large *British National Corpus* (BNC)). That is, students first engage with the words they are most likely to encounter in real life, and the words that they are most likely to misspell.

As students progress through the app, the vocabulary becomes more specialized and BNC exposure frequency tends to ‘flatten’. So, from the second 100-word tier onwards, words are also divided by the context where they are likely to be found. We use a loose categorization:

'Academic', 'Business', 'Common' and 'Technical' to help the student focus their efforts.

In each unit, the app then presents five sentences at a time, where each sentence contains a focus word with a gap, like a cloze exercise, where some letters have been removed. The app does *not* pronounce the word—the student is not learning dictation—but instead presents the word with crucial letters missing. The word itself is clear via the context of the complete sentence around it. The student has to touch the word and type the correct missing letters. For obvious reasons, the example sentences, all of which are drawn from a variety of corpora (see Sect. 7.2), were meticulously screened to ensure that they do not have any of the adjacent test words in them, even when shuffled and reshuffled while the student uses the app.

All three apps record the performance of the student on the device. They record the top score in a given exercise, and, if the clock timer is applied, the time taken to achieve that score. This means that students can see that their score has improved, they have become quicker and more adept at the task, or indeed have fallen out of practice. More competitive students may even challenge each other to beat their score or time by sharing the app.

The Spelling Practice module of ESP takes the concept of monitoring user performance to a new level. It consists of batteries of completion exercises (Fig. 8.5, left) that reproduce the student's spelling dilemmas, and a progress report (Fig. 8.5, right) that tracks the student's progress through the lexicon. Every time the student completes a particular word with a correct spelling, this fact is recorded. Each word will usually have between two and four *deletion patterns*, or variations of a single word with different letters removed. When a student correctly completes all deletion patterns for a word (or performs the task twice for rare cases with only one deletion pattern), the word is then recorded as 'completed'. Completed words are taken out of the exercise and placed in a list called *Words I Know* (Fig. 8.5, lower right), where they can be browsed and restored for more practice if required.

The effect of this procedure over time is that students who successfully complete the spelling of particular words will find that some exercise tests



Fig. 8.5 ESP's Spelling Practice module. The Progress Report (*right*) shows the user has completed 23 of the first 25 words, with a best score in that round of 100 per cent. The panel below lists the words learned in order of difficulty

have few words left to complete, and these are the words that they have found most difficult. To guide the student in deciding which words might require more practice, the 'Words I Know' list is ordered using the 'Wilson lower bound' scoring formula (Wallis 2013),⁴ which takes into account both the number of attempts and the number of errors that the student made. Progress is shown in a number of ways: in the exercise unit itself, on the menu listing all exercise units, and in the 'Words I Know' list.

⁴This method is commonly used to rank items by user-ratings on websites.

7 Wider Lessons for Linguistic Pedagogy Apps

What can these apps, and the design decisions that went into them, tell us about general principles for deploying language databases in pedagogical applications? In the following subsections we examine three different dimensions: the role of interactivity; the value of authentic language and grammatical analysis; and finally, the issue of how examples are appropriately selected from language sources.

7.1 Pedagogical Goals and the Interactivity Continuum

From the perspective of the programmer, mobile devices are small computers, which means that mobile app software can be developed on traditional computers and transferred to the device. Likewise, software initially developed for traditional computers (such as the *Internet Grammar of English* website) can in principle be redesigned to exploit the capabilities of such devices.

Technical discussion around apps is almost exclusively focused on *interface design*, because this is the first thing a developer must pay attention to. Thus a mobile phone will have a small screen and an imprecise pointing device (a finger) to control the interface. Any element that may be touched or dragged, from menus to exercise elements, must occupy an area on the screen surface somewhat larger than a fingertip. This places very different design constraints on the visual interface compared to the increasingly large screens and fine control offered by computer mice. A finger physically obscures small elements being touched, so a successful button press should generate feedback in some form (for example, sound, illumination around buttons, and so on).

The mobile platform has a number of other potential interface options, as well as constraints. These include vibration feedback, detecting orientation and rapid movement (shaking the device), accessing global positioning data, and access to mobile phone contacts, text messages and dialling. The three apps we have discussed are more conventional analogues

of desktop computer software, and with the exception of using vibration, do not try to exploit these capabilities.

Although the apps share common visual themes, interface, tools and structure, they are clearly distinguishable by where they sit on a spectrum of *interactivity*. By ‘interactivity’ we primarily mean the degree to which the user can usefully explore the app and its contents, rather than merely the degree to which *particular* content responds to user behaviour, which, thanks to a common codebase, is highly consistent across all the apps.

The least interactive app by this criterion is AWE, with its linear approach to the academic writing task, and only ten interactive exercises. To compensate a little, these exercises are supplemented by pop-up learning ‘tasks’ and checklists for authors. As we commented in Sect. 5, this lack of overall interactivity is not a drawback, given the pedagogical aims and provided that the user can still rapidly access relevant content.

At the other extreme, the app that depends the most on user interaction is ESP and its Spelling Practice module. Irregular English spelling is taught by a combination of exposure to real texts, and lots of practice. ESP offers the student the opportunity to practise spelling difficult words in exercises designed to be as close as possible to the cognitive task that the student undergoes when struggling to spell a word in a writing assignment. As discussed, the tracking of success for each individual word means that students are continually directed to the words that they find most difficult.

The first app to be developed, iGE, sits between these two extremes. It has some 36 different exercises, rather more than AWE. Indeed, ESP has fewer exercise types. But iGE does not depend pedagogically on exercise completion in the way that ESP does. Rather, the exercises in iGE aim to reinforce the course content and demonstrate the applicability of grammatical analysis to real language.

7.2 ‘Real Language’ and the Role of Grammar

All three apps draw on examples from corpora—whether for exemplification or exercise tests. The primary data set used for all apps is the parsed *British component of the International Corpus of English* (ICE-GB).

For ESP this data was supplemented by the Internet-generated corpus *WebCorp*,⁵ in order to provide citations of low-frequency words in use.

The interactive Grammar teaches a grammatical framework based on Quirk et al. (1985). The framework has been tested in the corpus, and the corpus provides the examples for the app. Since iGE is a course in grammar, the app focuses on grammatical terms throughout.

This framework is also employed in the other apps, although in these cases, grammatical concepts are only taught as needed by the pedagogical aims of the app. Thus in ESP, grammatical concepts are referred to in both regular morphology (for example, in discussing inflectional and derivational suffixes) and punctuation (for example, identifying subjects and objects, clauses, appositive noun phrases, and so on). Grammatical knowledge is not considered an end in itself in ESP.

Similarly, a small number of grammatical concepts are discussed in AWE's 'Style' chapter. These concepts, including prepositional phrases, phrasal verbs and ambiguous pronouns, are all sources of potential problems in writing for academic purposes. However, whereas the purpose of iGE is to teach a grammatical framework explicitly and completely, for AWE, grammatical *awareness*, rather than adherence to strictures, is more important. Using phrasal verbs appropriately in an academic register is part of the repertoire of a confident author. Being able to spot a phrasal verb is only part of the story: the skill is to identify a suitable Latinate replacement.

7.3 Corpora and the Problem of Example Selection

One of the most debated issues in language pedagogy is the relationship of language sources to the pedagogical process (see Sect. 2). In our apps, we have taken a pragmatic approach involving pre-selecting examples from running text in our corpora and creating a database of examples. These are then presented as 'authentic language' (which they are) in specific exercises and feedback is provided. But this is not the only approach that might be employed.

⁵ See <http://www.webcorp.org.uk>.

A very different approach is classroom concordancing (see Johns and King 1991), also known as ‘data-driven learning’. This is a type of exercise where students are encouraged to explore a computerized corpus for themselves, under the direction of a teacher. It has demonstrable benefits in certain areas, particularly translation studies, but it has been found not to be a universally successful language-learning approach (Kaltenböck and Mehlmauer-Larcher 2005). More specifically, classroom concordancing is *not optimal for every learning task*. Far from being truly autonomous, as is sometimes claimed, classroom concordancing relies on a high level of expertise by the teacher, who has to structure learning goals, direct a class, and crucially, provide expert feedback to students.

An analogy might be the use of microscopes in a classroom. Unless slides are preselected so that teachers can direct students’ attention to particular features, as well as anticipate students’ questions, a teacher will easily find herself or himself struggling to explain the presence of an unexpected and unfamiliar structure on any given slide.

A key question for pedagogical language tools is therefore *example selection*. We maintain that simply exposing students to a parsed corpus such as ICE-GB cannot substitute for a structured course in English grammar. First, students can be overwhelmed by too much information. Second, any given sentence beyond the most simple is likely to include grammatical constructions which are non-trivial for the teacher to explain. Like the biology teacher explaining an unexpected structure on a microscope slide, undirected learning of grammar is not for the faint-hearted. This is the old debate around ‘artificial’ and ‘contrived’ language for pedagogy (see Sect. 2) appearing as a technical issue of ‘database selection’.

As we have seen, the language-learning goals in our apps are varied. They include effective communication of critical thinking skills; high-level compositional conventions and awareness of audience expectations and register; learning grammatical terminology and analysis; and usage norms such as spelling and punctuation. Moreover, while these apps have been used in classrooms, they are also primarily designed for self-learning.⁶ This means that their design must allow the student to work

⁶However, see also *iGE in the classroom*, <http://www.ucl.ac.uk/english-usage/apps/ige/classroom.htm>.

through the material at their own pace, but they must be able to do so alone. Carefully calibrating the learning curve, and identifying how relevant assistance may be delivered, is therefore essential.

For most of the exercises in the apps, corpora offer a very large number of potential sentences from which to draw each test set. In ICE-GB alone, there are over 193,000 noun phrases headed by a noun. If we are to ask a student to identify noun phrases in example sentences, or pick out the head of the phrase, what principles should be used to select examples for the student to use? Are these principles capable of automation, so that, for example, a live corpus feed can generate examples ‘on the fly’? This latter possibility is not ideal for a self-contained app, where Internet connections are expected to be unreliable, but it is certainly feasible for web-based learning tools.

This selection problem must often be considered both individually and collectively. Individual criteria include relevance (for example, in an academic writing app, examples should be drawn from academic writing as far as possible) and readability of the test sentence. ‘Readability’ is a multi-faceted phenomenon, including *lexical complexity* (for example, ‘simplified measure of gobbledygook’, McLaughlin 1969), *lexical frequency* as an estimate of probable vocabulary familiarity (as in ESP), and *grammatical complexity* or estimates of processing load. Such automatic estimates of readability may be a useful guide.⁷ Still, whereas ‘grammatical complexity’ can seem like an intuitive measure of readability, another factor impeding reading fluency is the presence of unexpected grammatical structures that cause the reader to reread the sentence.

However, the selection problem is more complex than merely applying individual measures to text strings. In exercises where multiple examples are presented at one time, the problem must also be viewed holistically. For example, if a student is given five sentences and asked which are likely to be drawn from a text of ‘high’ or ‘low’ register, the vocabulary of both types of sentence should be generally comparable. Similarly, an exercise in iGE distinguishing between determiners and pronouns includes pairs of example sentences where the same lexical item appears in both examples (for example, *any ideas* vs *any of them*).

⁷ For child learning apps it may be appropriate to employ frequency statistics from age-appropriate corpora and also pre-screen the vocabulary in other ways.

An even more strict condition was applied to sentences in the Spelling Practice component of ESP. Example sentences were chosen so that none of the test words appeared in any form in any of the other sentences in the same exercise battery. For the test word *government*, neither *govern* nor *governing* could appear in any of the other 24 test sentences in the same battery.⁸

As we have also seen, exercises require ‘pedagogical scaffolding’, in the form of explicit explanation of principles and feedback. All three apps address this question by presenting interactive examples in exercise groups within a structured course. Although the student may jump straight into the exercise without reading the course material that precedes it, they are encouraged to follow the course. Terms are also linked back to the glossary. Moreover, as we saw, standardized feedback refers back to the explicit principles that they should have learned.

Many of the principles above are capable of automation. However, in the development of these three apps we focused on the viability of test examples and pedagogical aims, rather than relying on automatic methods. We therefore drew examples from corpora using corpus tools, and these examples went through a process of individual and collective evaluation against these criteria, before being finalized in the app.

8 Conclusions

The teaching of language, both to native and non-native speakers, has undergone massive changes in recent decades. Whereas artificial examples and textbooks used to be de rigueur, new technologies have opened up the possibility of novel ways for teaching language, and the most recent of these consists of smartphones and tablet computers.

The smartphone (or ‘handheld’) platform represents a new form factor for computing that has only become possible thanks to improvements in miniaturization, power consumption, efficiency and display screen

⁸This was further complicated after the first 100 words, as from this point on frequency differences between test words are rather smaller and less reliable. We therefore chose to additionally classify these words into four 25-word groups by their expected context (‘Academic’, ‘Business’ and so on).

technologies. This platform presents some significant interface and design challenges compared to the more traditional desktop. By contrast, the tablet platform is much closer to the desktop computer in screen size and selection precision. Our apps were designed first and foremost as smartphone apps, rather than tablet ones, as it was recognized that the smartphone represented the most significant design challenge.

We have shown how it is possible to draw examples from a corpus of natural language and apply them in these portable self-learning applications. This process was not automatic, although we elaborated some of the constraints that an automatic system would need to apply in selecting examples. The design of these apps revolves around their distinct pedagogical goals, rather than the technical capabilities of the devices on which they are to be addressed.

Finally, the aim of our apps is to improve students' knowledge of grammar and their writing ability. AWE in particular has been well received by academics as a helpful portable resource for improving students' writing and a weapon in the battle against student plagiarism.

References

- Aarts, Bas, Daniel Clayton, and Sean Wallis. 2012. Bridging the Grammar Gap: teaching English to the iPhone generation. *English Today* 28(1): 3–8.
- Aarts, Bas, and Sean Wallis. 2011. *interactive Grammar of English*. iOS and Android app. London: Survey of English Usage/UCL Business.
- Bax, Stephen. 2003. The end of CLT: a context approach to language teaching. *ELT Journal* 57(3): 278–287.
- Bloom, Benjamin, Max Engelhart, Edward Furst, Walter Hill, and David Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. New York: David McKay Company.
- Gass, Susan. 2003. Input and interaction. In *The Handbook of Second Language Acquisition*, eds. Catherine Doughty, and Michael Long, 224–255. Malden, Massachusetts: Blackwell.
- Gass, Susan, and Maria José Alvarez Torres. 2005. Attention when? An investigation of the ordering effect of input and interaction. *Studies in Second Language Acquisition* 27: 1–31.

- Gilmore, Alex. 2007. Authentic materials and authenticity in foreign language learning. *Language Teaching* 40(2): 97–119.
- Grabe, William. 2009. *Reading in a Second Language: Moving from Theory to Practice*. Cambridge: Cambridge University Press.
- Goodman, Kenneth, and David Freeman. 1993. What's simple in simplified language? In *Simplification: Theory and Application*, ed. Makhan Tickoo, 69–76. Singapore: SEAMEO Regional Language Center.
- Gunderson, Lee. 2009. *ESL (ELL) Literacy Instruction: A Guidebook to Theory and Practice*. New York: Routledge.
- Hudson, Richard, and John Walmsley. 2005. The English patient: English grammar and teaching in the twentieth century. *Journal of Linguistics* 41: 593–622.
- Johns, Tim and Philip King (eds). 1991. Classroom concordancing. *English Language Research Journal 4 (New Series)*. Birmingham University.
- Kaltenböck, Gunther, and Barbara Mehlmauer-Larcher. 2005. Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching. *ReCALL* 17(1): 65–84.
- Larsen-Freeman, Diane. 2000. *Techniques and Principles in Language Teaching*. Oxford: Oxford University Press.
- McCaul, Mae. 1973. Drills in language teaching. *TESL Reporter* 6(2): 6–7.
- McLaughlin, G. Harry. 1969. SMOG grading—A new readability formula. *Journal of Reading* 12(8): 639–646.
- Mehl, Seth, Sean Wallis, and Bas Aarts. 2013. *Academic Writing in English*. iOS and Android app. London: Survey of English Usage/UCL Business.
- Mitchell, Rosamond. 1994. The communicative approach to language teaching: an introduction. In *Teaching Modern Languages*, ed. Ann Swarbrick, 33–42. London: Routledge/The Open University.
- Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: Benjamins.
- Nesi, Hilary, and Sheena Gardner. 2012. *Genres across the Disciplines: Student Writing in Higher Education*. Cambridge: Cambridge University Press.
- Prodromou, Luke. 1996. Correspondence. *ELT Journal* 50(4): 371–373.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Royer, James, John A. Bates, and Christopher Konold. 1984. Learning from Text: methods of Affecting Reader Intent. In *Reading in a Foreign Language*, eds. Charles J. Alderson, and Alexander Urquhart, 65–81. London: Longman.

- Swan, Michael. 1985. A critical look at the Communicative Approach (1). *ELT Journal* 39(1): 2–12.
- Wallis, Sean. 2013. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics* 20(3): 178–208.
- Wallis, Sean, Seth Mehl, and Bas Aarts. 2014. *English Spelling and Punctuation*. iOS app. London: Survey of English Usage/UCL Business.
- Widdowson, Henry G. 1979. *Explorations in Applied Linguistics*. Oxford: Oxford University Press.
- 1985. Against dogma: a reply to Michael Swan. *ELT Journal* 39(3): 158–161.
- 1998. Context, community, and authentic language. *TESOL Quarterly* 32(4): 705–716.
- Willis, Dave. 1990. *The Lexical Syllabus: A New Approach to Language Teaching*. London: Collins.

Part II

Corpora for Continuing Professional Development

9

Locating People with Their Language: An Applied Linguistics Course Using Linguistic Microvariation Databases and Tools

Sjef Barbiers

1 Introduction

Recent years have seen a dramatic increase in the online availability of dialect corpora, databases and corresponding search, analysis and visualization tools for various dialect families such as Dutch, Scandinavian, Italian, English, Portuguese, Estonian and Slovenian.¹ Although primarily intended for linguistic research, this digital infrastructure also provides a rich resource for introductory and advanced courses on sociolinguistics, dialectology and the methodology of linguistic research. This chapter shows that the infrastructure is also very useful for courses in applied linguistics.

Three of the corpora in this online research infrastructure, DynaSAND (*Dynamic Syntactic Atlas of the Dutch Dialects*), GTRP (*Goeman*,

¹ See <http://www.dialectsyntax.org> for an overview and examples.

S. Barbiers (✉)
Meertens Instituut, Amsterdam, Netherlands
Utrecht University, Utrecht, Netherlands

Taeldeman, van Reenen Project) and DIDDD (*Diversity in Dutch DP Design*), all available in the online tool MIMORE (*Microcomparative Morphosyntactic Research* tool),² have been used in a course on Language Analysis for the Determination of Origin (LADO). Various countries use LADO in asylum procedures as one of the means to determine whether an asylum seeker was socialized in the country or area that s/he claims to originate from (see the papers in Zwaan et al. 2010). The determination of origin is crucial in the asylum procedure because only asylum seekers who have a well-founded fear of being persecuted in their country of origin for reasons of race, religion, nationality, political opinion or membership of a particular social group may get a permit to stay. LADO is used in particular when other ways to establish the origin have failed and there is serious doubt about the origin of the asylum seeker.

It is the responsibility of linguists to make clear if and how LADO can be a valid method. As is well-known from sociolinguistic research, the relation between a speaker's language properties and geographic origin is often highly problematic (see Patrick 2010). Various complications are demonstrated and discussed during the course, on the basis of the data in MIMORE, including the lack of reliable data, speaker-internal variation partially dependent on language background, accommodation dependent on speech situation, register and participants, unclear or gradual as opposed to categorical geographic distribution of certain linguistic features, the difference between linguistic and political borders and the requirement to make intuitions about geographic origin explicit qualitatively and quantitatively.

This chapter describes the role of MIMORE and its databases in the course *Locating People with their Language*. From 2011, this course has been taught several times as part of the Utrecht University Master's Program *Language, People and Society*. To make the role of MIMORE in this course transparent and put it in context, I provide a description of the content of the course, the way it was taught and how it was received by

²DynaSAND, <http://www.meertens.knaw.nl/sand>. GTRP, <http://www.meertens.knaw.nl/mand>. MIMORE, <http://www.meertens.knaw.nl/mimore>. MIMORE was developed by Jan Pieter Kunst, Matthijs Brouwer and Folkert de Vriend (Meertens Institute) in a CLARIN project (<http://www.clarin.nl>) led by Sjeff Barbiers.

the students. One clear finding is that students tend to overestimate their ability to derive geographic origin from language properties. Therefore, the knowledge provided in this course is essential. The course provides insight in the kind of linguistic expertise that is required to qualify as an expert LADO analyst. The usefulness of this course is not restricted to students of linguistics. Interpreters and language specialists (that is, native speakers) who play an important role in LADO procedures but usually do not have a background in linguistics would greatly benefit from this course as well, as a first step towards becoming an expert.

2 The MA Course: *Locating People with Their Language*

2.1 Introduction to the Course

The course starts with an introductory discussion of variation as an inherent property of natural language. The goal of this discussion is to replace commonplace ideas and myths about language variation with a scientific attitude that calls such ideas and myths into question. Fundamental questions are raised, such as:

1. Why is not there just one human language instead of the plethora of languages and lects that we find?
2. What could be the advantage of the human capacity to locate and identify people with their speech?
3. On the basis of which linguistic properties do we locate people in everyday life?
4. How reliable is this capacity?
5. What do we mean when we say that a person speaks Dutch and does this make sense?

We start with the distinction between language as a cognitive and as a social phenomenon. The students get some basic insight in the relation between language variation and factors such as gender, age, social class,

ethnicity, geographical origin. The effect of language variation is that a community (for example, a country or a region) is divided into groups which are partly overlapping and dynamic, giving rise to highly complex patterns. To a certain extent the group-defining capacity of language variation can be compared to that of clothing fashion.

The second question is whether such a division and the human capacity to recognize group membership has any advantage. To tentatively answer this question we look at language and humans from an evolutionary perspective. Human individuals live in groups and need groups to survive. It has been hypothesized that such groups have a maximal size of around 150 due to the limited social-cognitive capacities of human brains (see Dunbar 1998). This makes a division of communities into smaller groups necessary. Language variables function as shibboleths: it is important for humans to differentiate ingroups from outgroups.

The third question is meant to make the students aware that although we have clear intuitions about the group to which an individual belongs and hence about his or her geographic origin, it is very complicated to make explicit the linguistic properties on which such intuitions rely. It is clear that such explicitness should be required in LADO procedures, as it is in other legal procedures. Most intuitions rely on lexical and phonetic properties, which, however, are usually not categorical but gradual. There are also properties (in particular, syntactic properties) that are below the level of consciousness and therefore hard to observe and report for non-linguists. The reliability of our locating capacity (question 4) is dependent on these factors. Later in the course the students get an assignment to test the reliability of their own judgements (see Sect. 2.5).

As to question 5, the commonplace idea that the world can be divided into distinct dialects and languages is debunked. Language varieties usually cannot be demarcated very sharply, neither in the individual nor in society and space, and it makes more sense to speak of a continuum of language varieties. Therefore, it is often not possible to count languages and locate them in space in a sufficiently precise way, which is a serious complication for LADO procedures.

2.2 General Language Resources

In this part of the course the students are virtually placed in the position of a linguist who is responsible for a LADO procedure, as follows:³

Suppose there is an asylum seeker who claims to come from area A where people speak language or dialect B.⁴ If this claim is true, the asylum seeker will be granted a residence permit, because area A is dangerous (for him/her). The central part of the LADO procedure is an interview with the asylum seeker in his/her native language or in a second language that s/he speaks, carried out by an interviewer and an interpreter, both non-linguists. The recordings of the interview will then be analyzed by a native speaker, often a non-linguist, and you, the linguist who is responsible for the procedure and the final report. What kind of background information do you need to be able to carry out this procedure, which linguistic resources are available and how reliable are such resources?

A number of questions must be addressed in carrying out this procedure. Which languages and dialects are spoken where in the relevant area, when and by whom? What is the position of language B in this language situation? How (un-)stable is the language situation? What are the linguistic properties of language B? Which literature on this language is available? Are there linguistic atlases of the language area? Are there any linguistic specialists on this language? Are any of them native speakers of it?

The students are asked to explore general language resources (mainly on the web) that (may) play a role in LADO procedures. The resources used include, among others, *Ethnologue*, *UNESCO Atlas of the World's Languages in Danger*, *The World Atlas of Language Structures Online*, *Eurominority* and *LinguistList*.⁵ The students must write and present a

³This is just one type of LADO procedure. A range of types exists, for example LADO procedures in which the interviewer is a linguist or uses no interpreter.

⁴In the remainder of this chapter I will use the term 'language' to refer to both languages and dialects, except in cases where it is necessary to distinguish between the two.

⁵For further details of these resources, see: *Ethnologue* (<http://www.ethnologue.com>), *UNESCO Atlas of the World's Languages in Danger* (<http://www.unesco.org/languages-atlas>), *The World Atlas of Language Structures Online* (<http://wals.info>), *Eurominority* (<http://www.eurominority.eu>) and *LinguistList* (<http://www.linguistlist.org>).

report on which types of information these resources make available, and they also give an evaluation of the reliability of this information.

The latter is very important because quite a few students tend to take the reliability of such information for granted, especially if it comes from official sources such as SIL (originally known as the Summer Institute of Linguistics, the organization behind *Ethnologue*), UNESCO, Max Planck Institute (WALS) and so on. They find out that the number of linguistic atlases that can be useful in LADO procedures is very limited. They gain the fundamental insight that language situations are constantly changing and therefore that information on the number of languages in an area, number of speakers, location of the language, language properties and so on can never be complete and fully up to date.

If a resource nevertheless provides such information, checks need to be made on when the data on which this information is based were collected, how they were collected and by whom. For example, it makes a huge difference whether the data come from a recent census or reflect the intuitions of an individual linguist. It is also important to take into account the potential negative impact of self-reporting on the reliability of the data, especially self-reports by bilingual and bilialectal speakers. As an example, we identify in the course, on the basis of the MIMORE data, a set of speakers in the western part of the Netherlands who consider themselves speakers of Standard Dutch but clearly are (also) speakers of a Hollandic dialect. Such speakers may deny the existence of a certain linguistic property in their dialect while at the same time using it in spontaneous speech.

Eventually, the students should be able to judge the extent to which a piece of information from such linguistic resources can be reliably used in the LADO report on the asylum seeker from area A who claims to speak language B. They should also understand the risks and consequences of the choice of a particular interpreter and language analyst in the framework of LADO. It is important to know whether these participants in the LADO procedure speak the same variety as the asylum seeker or a language variety related to it, and if the latter, how these varieties differ from each other.

2.3 Sociolinguistic Variation and Multilingualism

This part of the course addresses the problem of sociolinguistically determined variation and multilingualism, on the basis of Patrick (2010) and Muysken (2010), two papers that the students have to read. The goal is to understand language variability in its full complexity and to learn to relate this complexity to the possibility of determining the origin of a speaker on the basis of his/her language. I will give a summary of the insights presented in these two papers that play a central role in the course.

A key issue is whether the language background of a specific speaker allows us to determine his/her origin. A distinction needs to be made between geographic area and speech community as the origin of a speaker. There are various situations in which these two do not coincide. This is for example the case when a speaker is a member of an immigrant group that have retained their native language and not adopted a language of the receiving area. With LADO it may well be possible to determine the speech community from which the speaker originates, but impossible to determine his or her geographic origin. This is a serious complication, because it may very well be the situation in his/her geographic area, rather than the speech community, that brings the asylum seeker to the decision to flee. Another example of such a situation is when a speech community is spread over a geographical area that crosses a political border, with a safe and a dangerous side.

The language variety or varieties that a person speaks depend on a large array of factors, including the language(s) of the parents, the languages in the environment (especially of peers), age, sex, social class, ethnicity, education. This means that a speech community is never homogeneous. This raises the question of which variety has to be taken as the standard of comparison when trying to establish the speech community from which the speaker originates.

The mobility of the speaker and his/her family is another crucial factor. When a speaker has a history of moving from place to place, especially during childhood, s/he may speak a range of language varieties. These varieties may influence each other, with the result that in a concrete speech situation the speaker may switch between the language varieties

and his/her proficiency may vary from variety to variety (Muysken 2010). The linguist responsible for a LADO procedure therefore needs to investigate the language background of the speaker to establish whether s/he can be subjected to such a process.

Furthermore, the language use of a speaker is known to depend on the speech situation, a crucial factor for LADO interviews which of course constitute highly artificial and formal situations (see the description of the interview situation in Sect. 2.2). Situational factors determining language properties include the bureaucratic context, unequal power relations among the interview participants, necessity for language choice and interpretation, existence of ethnic, class or racial conflicts which affect cross-cultural communication, pressures on minorities to assimilate linguistically to majorities, and a tendency to accommodate to the language that has a higher prestige, prevalence of language contact, code-switching and language mixing, and prescriptive language attitudes and ideologies (Patrick 2010: 79).

Another crucial insight that the students need to gain is that linguistic differences between language varieties are often not categorical but gradual and need to be measured both qualitatively and quantitatively. When students are asked to determine the geographic origin of a speech fragment, they typically motivate their answers with: the speaker is using word X which is typical for the south, or, the speaker does not pronounce sound Y, which is typical for the east. Students tend to take such properties as categorical, while they often are not. Furthermore, whether a particular sound is pronounced or not also depends on the linguistic context within a language variety, for example sound Y is pronounced in stressed syllables but not in unstressed ones. Also, two language varieties may differ in the number of times that a sound is pronounced, all things being equal. For example, a speaker of southern Dutch may pronounce /t/ in /niet/ 'not' 20 per cent of the time, while a speaker of central Dutch does that 80 per cent of the time. Therefore, the students have to learn to analyse speech fragments and their transcriptions linguistically, both qualitatively and quantitatively.⁶

⁶Note that such details of frequency are almost never available in LADO reports as they are unknown for the target languages.

2.4 Microvariation Databases, Tools and Atlases for Language Analysis

2.4.1 Introduction

In this part of the course the students get acquainted with the basics of linguistic analysis necessary for locating speech fragments, using a number of Dutch resources. There are several reasons to choose resources on varieties of Dutch, despite the fact that LADO in the Dutch situation usually involves language varieties spoken in parts of Africa and Asia.

First of all, the students in this course are Dutch themselves and hardly ever speak or understand one of the languages relevant for LADO in the Netherlands. It would therefore be very complicated to demonstrate the intricacies of linguistic intuitions and analysis with such foreign language varieties.

A second reason for choosing varieties of Dutch is that this part of the course has an additional goal to let students experience how difficult it is to make linguistically explicit an intuition about the geographic origin of a language variety related to their own. Varieties of Dutch serve this goal well. When a student who speaks Standard Dutch hears a speech fragment of a Dutch dialect variety for the first time s/he will have certain intuitions about the origin. S/he will then be asked to make these intuitions explicit by comparing the linguistic properties of this variety with those of Standard Dutch. Not only will the student experience how difficult this is, but also, the student is in this way in the same position as the (non-linguist) language specialist in a LADO procedure who is the speaker of a national language and has to judge a dialect of that language. Put differently, the student will get a first idea of whether a language specialist in such a situation will be able to come to valid conclusions and motivate them.

A third reason is that in this part of the course the students need to learn that certain linguistic properties are below their level of consciousness. They will not notice certain differences between their Standard Dutch variety and a Dutch dialect until the teacher has made them aware of these. This is particularly the case for (morpho-)syntactic differences. When asked to mention some (morpho-)syntactic differences between

Standard Dutch and familiar Dutch dialects, students are generally not able to do so, while they are able to give examples of differences at the lexical or phonetic level. Similarly, when asked to give the linguistic differences between a particular speech fragment of a Dutch dialect and Standard Dutch, they typically overlook the (morpho-)syntactic differences.

For example, there are Dutch dialects in which, in addition to the finite verb, the complementizer agrees with a plural subject, with a /-n/ or /-ə/ suffix as the exponent. Such minimal (morpho-)syntactic differences usually go unnoticed, not only by the speakers of the relevant varieties themselves (see Pauwels 1958 and Barbiers 2015) but also by speakers of related varieties that do not have these properties. (Morpho-)syntactic properties therefore are potentially determining factors in a LADO procedure, as a speaker trying to imitate a variety will typically omit (morpho-)syntactic properties that are below the level of consciousness.

To teach students the relevance of minimal, often subconscious linguistic differences and how to recognize and describe them, we introduce them to a number of Dutch microvariation databases and corresponding software tools. The software tool MIMORE gives access to three databases: the GTRP database on phonological and morphological variation in the Dutch language area, the DynaSAND database on syntactic and morphosyntactic variation at the clausal level, and the DIDDD database on syntactic and morphosyntactic variation at the level of the nominal group.⁷ The students use these three databases for their assignments. There now follows a brief description of the content and functionality of GTRP, DynaSAND, DIDDD and MIMORE.

2.4.2 GTRP

The GTRP database includes the results of a data collection project carried out between 1979 and 2000 under the responsibility of the linguists Goeman, Taeldeman and van Reenen.⁸ Data were collected in 611 locations across the Netherlands (including Frisia), Flanders (that is, the

⁷ See <http://www.meertens.knaw.nl/mimore>.

⁸ Users who understand Dutch can access the GTRP database online at <http://www.meertens.knaw.nl/mand/database>. Other users should use the MIMORE tool (see Sect. 2.4.5).

Dutch-speaking part of Belgium) and French Flanders, a small part of north-west France. The informants in these locations were asked to translate a list of 1876 items. These items were mainly individual words and phrases, and sometimes complete sentences. All informants had to meet the following requirements:

- The informant speaks the dialect of the community;
- The informant is born in the place of residence and has lived there preferably his/her whole life; the same goes for his/her parents;
- The informant is between 50 and 75 years old;
- The informant is preferably low-educated but with considerably high literacy skills.

Given the goal of the project, to chart phonological and morphological variation in the Dutch language area, these informant requirements were necessary to ensure that there was a relation between linguistic variable and geographic location and to reduce the potential influence of other sociolinguistic variables such as social class and age.

The atlases resulting from this project include the phonological atlas of the Dutch dialects FAND (three volumes; FAND I: Goossens et al. 1998; FAND II + III: Goossens et al. 2000; FAND IV: De Wulf et al. 2005) and the morphological atlas of the Dutch dialects MAND (two volumes: MAND I: de Schutter et al. 2005 and MAND II: Goeman et al. 2008). Together the three FAND volumes give a detailed overview of variation in the vowel and consonant systems of the Dutch dialects and the geographic distribution of this variation. The two MAND volumes give an overview of the variation in plural formation, diminutives, gender, comparatives and superlatives, possessive pronouns, subject and object pronouns, verbal inflection, participles and verb stem alternations.

2.4.3 DynaSAND

For an extensive description of DynaSAND and its background, see Barbiers et al. (2007) and Barbiers and Bennis (2007).⁹ The data in DynaSAND were collected between 2000 and 2003 in 267 locations in

⁹ DynaSAND can be accessed at <http://www.meertens.knaw.nl/sand> and in MIMORE.

the Netherlands, Flanders and north-west France. The basis of the selection of locations was an even distribution across the language area, with higher density in areas where the dialects are still very strong and numerous, in transitional zones and in locations with special circumstances, for example (former) islands. The goal of the project was to chart the geographic distribution of (morpho-)syntactic variation at the clausal level. Therefore the informants in the fieldwork stage had to meet more or less similar requirements to the GTRP informants.

The methodology of data collection for DynaSAND was different from GTRP, though. There were three stages: a postal pilot study, oral interviews and telephone interviews. The atlases (SAND I, Barbiers et al. 2005; SAND II, Barbiers et al. 2008) are based on the oral and the telephone interviews. In the oral interviews in the Netherlands, there were two informants in each location and they did the interview together in the local dialect without interventions by the fieldworker. This was to reduce accommodation as much as possible and to avoid judgements based on phonetic and lexical differences. As opposed to the Dutch interviews, the Flemish interviews were carried out by linguists who spoke the dialect or regiolect of the area and there were two informants in each location.

Around 150 different syntactic properties in 424 test sentences were investigated. We mainly used translation tasks and concealed judgement tasks. The latter did not ask for the grammaticality but for the commonality of a construction in a particular dialect, to avoid influence of normativity on the judgements. Often translation and judgement tasks were combined, also to check whether translation and judgement were consistent. Other types of tasks that we used include cloze tests, completion tasks and picture response tasks.

The data available in DynaSAND and the two SAND volumes include the left periphery (complementizer system, complementizer agreement, Wh questions, relative clauses, other fronting constructions), subject pronouns, subject pronoun doubling and cliticization, reflexive and reciprocal pronouns, morphosyntax of verbal clusters and auxiliaries, verb cluster interruption, negation and quantification.

With the DynaSAND software tool it is possible to search the database with text strings, strings of parts-of-speech (POS) tags, test sentences, syntactic phenomena, locations and areas. The results of these searches

are lists of geo-referenced sentences with tagging, English glosses and translations and the corresponding sound fragments. This makes it possible to check the validity of the data and to select the locations that have a particular syntactic phenomenon. This selection can then be fed into a cartographic tool that depicts the geographic distribution of the phenomenon and in the case of multiple phenomena, the correlations between them. Most of the maps of the printed volumes SAND I and II are also available in DynaSAND by searching with syntactic phenomena.

2.4.4 DIDDD

The data for the *Diversity in Dutch DP Design* database were collected between 2005 and 2009 in about 200 locations in the Netherlands, often the same locations as in DynaSAND and with a methodology comparable to that of DynaSAND. For a more extensive description see Corver et al. (2007). The DIDDD data can be accessed through MIMORE.

The goal of DIDDD was to describe the variation in nominal groups in the Dutch dialects. Phenomena investigated include noun phrase internal pronouns, substantivized pronouns, combinations of definite articles, demonstratives and possessive pronouns/phrases, number, negation, and quantification. For an overview of attested variation see Corver et al. (2013).

2.4.5 MIMORE

MIMORE was developed to enable the researcher to search DynaSAND, GTRP and DIDDD at once and in a uniform way.¹⁰ In this way morphological properties, (morpho-)syntactic properties at the level of the nominal group and (morpho-)syntactic properties at the clausal level can be related to each other. It is possible to search with text strings, strings of POS tags and syntactic phenomena (Fig. 9.1).

¹⁰ See <http://www.meertens.knaw.nl/mimore>.



Fig. 9.1 The MIMORE search tool

The POS tags to be included in the search can be constructed from a list of primitive categories and a list of primitive features, or one can use a list of predefined complex tags.

As in the case of DynaSAND, the result of a search is a list of geo-referenced sentences or sentence fragments with their POS taggings, glosses and translations and the corresponding sound clips (if available). In the case of search results from GTR, phonetic representations are given as well, which is relevant for students who need to compare pronunciation in detail. Selections of search results can be exported. One way of exporting is to a so-called virtual collection which makes further operations possible. It is possible to derive the intersection, union and complement set of sets of selected search results (that is, of their locations). This way, potential correlations between two or more phenomena can be investigated. Sets of search results can also be presented on geographic maps (Fig. 9.2).

MIMORE Virtual Collection

perform new search 2 items in your virtual collection

#	nr		Title	Description	Items	Created	
<input type="checkbox"/>	1	 	Syntactic Phenomenon	256 results from search for syntactic phenom 'Complementizer agreement second person singular: <math>\langle \rangle \text{dat-} \text{st} \langle \rangle \text{/} \rangle \text{' in sand No syntactic phenomena specified to search for in gtrp, resource ignored! No syntactic phenomena specified to search for in diddd, resource ignored!	256	29/11/2015, 00:33:59	
<input type="checkbox"/>	2	 	Syntactic Phenomenon	787 results from search for syntactic phenom 'Second person singular, after V' in sand No syntactic phenomena specified to search for in gtrp, resource ignored! No syntactic phenomena specified to search for in diddd, resource ignored!	787	29/11/2015, 00:40:04	

group by location in maps show frequency in maps

Select at least one item to enable the additional functionality on sets of items like e.g. combined maps and copy & merge

© 2000-2015 KNAW / Meertens Instituut

MEERTENS INSTITUUT Postbus 94264, 1090 GG Amsterdam. Telephone +31 (0)20 4628500. Fax +31 (0)20 4628555. info@meertens.knaw.nl

Fig. 9.2 MIMORE virtual collection

2.5 Use of Microvariation Databases and Tools in the Course

After having read some introductions to the databases, tools and atlases involved (Barbiers 2006; Goeman 2006; Taeldeman 2006; Corver et al. 2007), the students get an assignment developed by Wilbert Heeringa (Groningen University). It involves speech fragments with identical content from eight different locations/dialects in the Netherlands and Flanders. Six of these locations are close to the Dutch–Belgian national border; the fragments of the two other locations serve as control items. The main question of this assignment is: is it possible to determine whether a certain fragment is from the Belgian part of the language area or from the Netherlandic part? This mimics the situation in a LADO procedure, in which an informant must be located in the right area, where ‘right area’

is politically defined. In this assignment, we are dealing with a political division of one language area (the Dutch one) and the dialects on both sides of the border are closely related—we have three cross-border minimal pairs of Limburgian, Brabantish and Flemish dialects.

The assignment consists of the following steps. First, the students listen to the eight fragments. Then they are asked to transcribe (some of) the fragments orthographically and partly in IPA. Using these transcriptions, they give a detailed description of the lexical, phonetic, morphological and syntactic differences between each fragment and Standard Dutch.¹¹ Finally, they have to determine on the basis of these descriptions whether a speech fragment belongs to the Netherlandic or the Belgian part of the language area and where in the area the speech fragment is from. To be able to make these final steps they need to compare their descriptions of the sound fragments with the transcriptions and sound fragments that are available in the three databases in MIMORE and with the maps in the various printed atlases.

There are various ways in which the students can use the databases and tools. A first option is to search for location. This is useful in cases where the student has an intuition about where a speech fragment could be located but does not yet know how to support this intuition with linguistic properties. In such a case s/he can search, for example, in DynaSAND with one or more location codes or names or with a geographic region or province. The results of such a search are the complete interviews for one or more locations. The student can then start to compare the transcriptions and sound recordings of these interviews with the material to be located.

A second way of using the databases and tools obtains when the student has no idea where a speech fragment should be located, but s/he has found a number of linguistic properties that are distinct from Standard Dutch. In this situation, s/he can start searching with these linguistic properties, with text strings, POS tag strings, test sentences and syntactic phenomena, to find out in which locations or areas these properties occur.

Obviously, these two methods can and should be combined. For example, when a student uses the second method and has found a number

¹¹ Comparing dialect and standard language is not a typical part of LADO tasks but it is necessary to train the students to identify and describe linguistic features.

of locations, the next step should be to go to the full interviews of these locations and make a detailed comparison between these interviews and the fragment to be located in order to find a list of common properties. Again, it is very important in this stage of the course that the student learns that it is not enough to say: I found this word/sound/construction, therefore the fragment must from location/area X. Rather, the outcome of this assignment should be that the student understands that only a list of (preferably quantified) properties from various linguistic domains will be convincing evidence for a particular location.¹²

The results of this assignment in the past four years show that students tend to stick to the lexical and phonetic level when describing the properties of a dialect. This is a striking result given the enormous amount of syntactic and morphological data available, and the fact that syntactic and morphological properties are often more distinctive and decisive than lexical and phonetic properties. As was suggested in Sect. 2.4.1, they possibly neglect syntactic and morphological properties because such properties are often subconscious. The next question is of course why that would be the case. Needless to say, as soon as the students have been made aware of particular syntactic and morphological properties that are typical for dialects, their performance on this decision task improves. Another finding is that there seems to be a discrepancy between the ability to decide where in the language area a fragment should be located and the ability to make such a decision linguistically explicit. That is, the average student performs not flawlessly but reasonably well on the location task, but there is quite some variation in the quality of linguistic argumentation that is used to reach such a decision.

2.6 A Location Experiment

Following up on the course described here and inspired by Foulkes and Wilson (2011), Smidt van Gelder designed an experiment to find an answer to the question: can analysts correctly locate speakers whose

¹²Note that LADO analysis normally begins with a pre-existing set of features that distinguish the targeted dialects, rather than generating a list of features that strike the analyst as salient while listening, as described here.

dialect is not the same as the analyst's, but a related one; for example, can speakers of British English correctly locate speakers of American English dialects? The experiment and its results are reported in Smidt van Gelder (2012). The question she asked is directly relevant for the LADO procedure. Recall that it is not always possible to find a native speaker of exactly the same dialect for the roles of language analyst and interpreter and that in such cases speakers of related language varieties will be chosen. It is important to know how this influences the validity of the conclusion.

Smidt van Gelder presented a number of speech fragments of Limburgian dialects close to the Dutch–Belgian border to groups of (non-linguist) subjects at increasing distances from the geographic origin of the speech fragment. Unsurprisingly, she found that the further the distance, the more mistakes in locating the speech fragment. However, she also found that even speakers of exactly the same local dialect make mistakes and are not always able to determine whether a speaker is from his/her own village or from a neighbouring village close by across the border.

Smidt van Gelder concludes that the speakers in her experiment were not very good at locating a speaker of a related dialect, that their performance got worse with increasing geographic/linguistic distances, that they were very bad in motivating their decisions, that these motivations were usually superficial and based on feelings rather than facts and that despite all of this the speakers were very confident that their locating judgements were right.

The question should therefore be raised as to whether native speakers who do not speak exactly the same language variety as the asylum seeker should play a role in LADO procedures. This research also shows that the people involved in LADO procedures should be trained in linguistics to be able to use and analyse linguistic resources and to provide explicit and factual motivation when they try to locate a person.

2.7 Remainder of the Course

For the sake of completeness, I describe the final part of the course, in which MIMORE and its databases play a more limited role. Up to this point, the course has introduced the students to the availability and use of linguistic resources, the basics of linguistic variation, the LADO procedure

and the problems involved in it. In the next stage of the course, the students attend three lectures by guest speakers.

The first guest speaker is a linguist who is responsible for LADO procedures at the Dutch immigration and naturalization service.¹³ She discusses the various aspects and steps of the procedure. Central in this lecture is the question of whether it is possible to locate a speaker on the basis of his or her *second* language. More specifically, is it possible to differentiate the L2 English of speakers from various parts of Africa, to make the differences between these L2 varieties of English explicit and relate them to the L1s of the speakers? With Simo Bobda et al. (1999) as background reading, the students have to analyse a speech fragment from an African English speaking informant and to derive the geographical origin of this speaker, using online databases of English accents and other resources.

The second guest speaker is a linguist who works for the Taalstudio, a Dutch company that makes (among others) counter-expert reports for asylum seekers in LADO procedures.¹⁴ She discusses the debate on whether native speakers can be reliable language analysts in LADO (see Cambier-Langeveld 2010), the guidelines for LADO (see the annex in Zwaan et al. 2010) and the requirements for a scientifically responsible approach to LADO and the research needed for that (see McNamara et al. 2010).

LADO can be considered as a forensic application of linguistics. To make the students familiar with other forensic applications of linguistics there is a third speaker, a phonetician who works for the Dutch forensic institute.¹⁵ He discusses the role of language analysis in criminal investigations, not only to locate people but also to derive a speaker profile and to establish whether two speech fragments recorded on two distinct occasions can be traced to the same speaker. Here the data in MIMORE become relevant again to make judgements on speech fragment origin, identity and speaker's profile explicit.

For the final session of the course, the students have to analyse a public debate on the racist anti-Islam pamphlet *De ondergang van Nederland, land der naieve dwazen* (The decline of the Netherlands, country of

¹³ See <https://www.ind.nl>.

¹⁴ See <http://www.taalstudio.nl>.

¹⁵ See <http://www.forensischinstituut.nl>.

naive fools). This pamphlet was published in 1990 under the name of Mohamed Rasoel, and the public debate is about whether the Dutch author Gerrit Komrij could be the real author. The students have to evaluate the argumentation in this debate, which is mainly about style. They get a brief introduction in measuring style and the use of stylistic analysis tools, such as the demonstrator *Stylene* of the University of Antwerp.¹⁶

3 Conclusion

From the course evaluations it is clear that this course opens up a whole new world for the students, not only the world of linguistics but also the relevance of linguistics for problems in the real world. They learn to look at language variation in a new way, ask fundamental scientific questions about it, become familiar with resources that show the actual variation, and learn to question and evaluate the validity of such resources. They then learn to analyse real language fragments, using the online tools, databases and maps available in MIMORE. It becomes clear that such tools and resources are indispensable for a serious attempt to locate a speech fragment and its speaker. Unfortunately, the Dutch language area is one of the few language areas in the world for which such extensive and detailed resources are available.

References

- Barbiers, Sjef. 2006. De Syntactische Atlas van de Nederlandse Dialecten. *Taal & Tongval* 18: 7–40.
- . 2015. European Dialect Syntax: Towards an infrastructure for documentation and research of endangered dialects. In *Endangered Languages and New Technologies*, ed. Mari C. Jones, 35–48. Cambridge: Cambridge University Press.
- Barbiers, Sjef, Hans J. Bennis, Gunther de Vogelaer, Magda Devos, and Margreet H. van der Ham. 2005. *Syntactische Atlas van de Nederlandse Dialecten/Syntactic Atlas of the Dutch Dialects*, vol I. Amsterdam: Amsterdam University Press.

¹⁶ See <http://www.clips.ua.ac.be/cgi-bin/stylenedemo.html>.

- Barbiers, Sjef, and Hans J. Bennis. 2007. The Syntactic Atlas of the Dutch Dialects: A discussion of choices in the SAND-project. *Nordlyd* 34: 53–72.
- Barbiers, Sjef, Leonie Cornips, and Jan Pieter Kunst. 2007. The Syntactic Atlas of the Dutch Dialects: A corpus of elicited speech and text as an on-line dynamic atlas. In *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*, eds. Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, 54–90. Basingstoke: Palgrave Macmillan.
- Barbiers, Sjef, Johan van der Auwera, Hans J. Bennis, Eefje Boef, Gunther De Vogelaer, and Margreet H. van der Ham. 2008. *Syntactische Atlas van de Nederlandse Dialecten Deel II/Syntactic Atlas of the Dutch Dialects*, vol II. Amsterdam: Amsterdam University Press.
- Cambier-Langeveld, Tina. 2010. The validity of language analysis in the Netherlands. In *Language and Origin: The Role of Language in European Asylum Procedures: Linguistic and Legal Perspectives*, eds. Karin Zwaan, Maaïke Verrips, and Pieter Muysken, 21–34. Nijmegen: Wolf Legal Publishers.
- Corver, Norbert, Marjo van Koppen, Huib Kranendonk, and Mirjam Rigterink. 2007. The noun phrase: Diversity in Dutch DP design (DiDDD). *Scandinavian Dialect Syntax* 34: 73–85.
- Corver, Norbert, Marjo van Koppen, and Huib Kranendonk. 2013. De nominale woordgroep vanuit dialect-vergelijkend perspectief: Variaties en generalisaties. *Nederlandse taalkunde* 18(2): 107–138.
- De Schutter, Georges, Boudewijn L. van den Berg, Ton Goeman, and Thera de Jong. 2005. *MAND Morfologische Atlas van de Nederlandse Dialecten Deel II/ MAND Morphological Atlas of the Dutch Dialects*, vol I. Amsterdam: Amsterdam University Press.
- De Wulf, Chris, Jan Goossens, and Johan Taeldeman. 2005. *Fonologische Atlas van de Nederlandse Dialecten. Deel IV. De consonanten*. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.
- Dunbar, Robin I.M. 1998. The social brain hypothesis. *Evolutionary Anthropology* 6(5): 178–190.
- Foulkes, Paul, and Kim Wilson. 2011. Language analysis for the determination of origin: An empirical study. *Proceedings of the 17th International Congress of Phonetic Sciences* 17: 691–694.
- Goeman, Ton. 2006. De Morfologische Atlas van de Nederlandse Dialecten (MAND); zero en bewaard gebleven morfologische informatie. *Taal & Tongval Themanummer* 18: 66–92.
- Goeman, Ton, Marc van Oostendorp, Piet van Reenen, Oele Koornwinder, Boudewijn L. van den Berg, and Anke van Reenen. 2008. *Morfologische atlas*

- van de Nederlandse dialecten. Deel II/Morphological Atlas of the Dutch Dialects*, vol II. Amsterdam: Amsterdam University Press.
- Goossens, Jan, Johan Taeldeman, and Geert Verleyen. 1998. *Fonologische Atlas van de Nederlandse Dialecten. Deel I*. Gent: Koninklijke Academie voor Nederlandse Taal en letterkunde.
- Goossens, Jan, Johan Taeldeman, and Geert Verleijen. 2000. *Fonologische Atlas van de Nederlandse Dialecten. Deel II. De Westgermaanse korte vocalen in open lettergreep. Deel III. De Westgermaanse lange vocalen en diftongen*. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.
- McNamara, Tim, Carolien van den Hazelkamp, and Maaïke Verrips. 2010. LADO, validity and language testing. In *Language and Origin: The Role of Language in European Asylum Procedures: Linguistic and Legal Perspectives*, eds. Karin Zwaan, Maaïke Verrips, and Pieter Muysken, 61–72. Nijmegen: Wolf Legal Publishers.
- Muysken, Pieter. 2010. Multilingualism and LADO. In *Language and Origin: The Role of Language in European Asylum Procedures: Linguistic and Legal Perspectives*, eds. Karin Zwaan, Maaïke Verrips, and Pieter Muysken, 89–98. Nijmegen: Wolf Legal Publishers.
- Patrick, Peter L. 2010. Language variation and LADO. In *Language and Origin: The Role of Language in European Asylum Procedures: Linguistic and Legal Perspectives*, eds. Karin Zwaan, Maaïke Verrips, and Pieter Muysken, 73–88. Nijmegen: Wolf Legal Publishers.
- Pauwels, Jan L.H. 1958. *Het dialect van Aarschot en omstreken*. Brussels: Belgisch Interuniversitair Centrum voor Neerlandistiek.
- Bobda, Simo, Hans-Georg Wolf Augustin, and Lothar Peter. 1999. Identifying regional and national origin of English-speaking Africans seeking asylum in Germany. *Forensic Linguistics* 6: 300–319.
- Smidt van Gelder, Nadia. 2012. Wao kûmst dich vanaâf? Taalanalyse in asielpcedures: het vermogen van native speakers om sprekers van een ander dialect te herkennen. Unpublished MA thesis. Utrecht University.
- Taeldeman, Johan. 2006. De Fonologische Atlas van de Nederlandse dialecten (FAND). Opzet, uitwerking en operationaliteit. *Taal & Tongval Themanummer* 18: 116–147.
- Zwaan, Karin, Maaïke Verrips, and Pieter Muysken, eds. 2010. *Language and Origin: The Role of Language in European Asylum Procedures: Linguistic and Legal Perspectives*. Nijmegen: Wolf Legal Publishers.

10

From Sociolinguistic Research to English Language Teaching

Jenny Cheshire and Susan Fox

1 Introduction: The Educational Context

Sociolinguistic research on spoken English and language variation has acquired great importance in the UK secondary school curriculum with the introduction in 1981 of the GCE (General Certificate of Education) Advanced level examination in English Language. GCE examinations are taken by school leavers at the age of 18 in England, Wales and Northern Ireland; they are also taken by students in Scotland as an alternative to Advanced Higher examinations and as an international qualification around the world. In the UK the grades obtained determine whether students are accepted for university education, with the best universities requiring very high grades. Generally, students take GCE examinations in three or four subjects of their choice.

J. Cheshire
Queen Mary University of London, London, UK

S. Fox (✉)
University of Bern, Bern, Switzerland

GCE English Language has been steadily increasing in popularity since it was introduced and it is now one of the fastest-growing school subjects. During the period 2003–10, for example, the number of entries for all five examining boards rose from 14,751 in 2003 to 23,211 in 2010.¹ Teachers of this subject are required to ‘introduce students to the concepts and methods of the disciplines of English language/linguistics in relation to a wide range of spoken and written forms of English, including electronic and multimodal forms’ (Department for Education 2014: 1). They must accurately use a range of terminology associated with phonetics, phonology, prosody, lexis and semantics, ‘grammar including morphology’, ‘pragmatics and discourse’ and show how the terminology can be applied to a range of contexts for language use, including historical, geographical, social and individual varieties of English, and aspects of language and identity (Department for Education 2014: 2).²

While we may applaud the government’s drive to educate young people to use linguistic concepts and terminology to accurately analyse language in use, in practice the specifications place heavy demands on teachers. Most English schoolteachers have a background in English literature, not language—two very different disciplines—and so have little or no training in linguistic analysis, especially as applied to spoken rather than written language. In-service training is unavailable in many areas of the UK, so teachers have to learn new subject knowledge largely on their own (Bleiman and Webster 2006: 29). As a result, many teachers are ‘seriously lacking in confidence’ and ‘often feel overwhelmed and uncer-

¹ See www.phon.ucl.ac.uk/home/dick/ec/stats.htm.

² At the time of our project, in 2011, the specifications for this examination incorporated still more sociolinguistics: for example, a section on Language Variation and Change included the study of standard and vernacular dialects and accents, and debates about the role of standard and vernacular varieties in education. For a brief period, between 2010 and 2014, the GCSE (General Certificate of Secondary Education) English Language and English Literature examinations also included the study of spoken English, with the English Language specifications including the study of variation in spoken English and its relation to identity and cultural diversity. This examination is taken at age 16; students choose between 1 and 10 subjects, with English Language or English Literature compulsory. In 2014 the Secretary of State for Education, Michael Gove, changed the GCSE curriculum so that it now focuses almost entirely on written English, with a small and unassessed component dealing with using ‘spoken standard English effectively in speeches and presentations’ (Department for Education 2013: 6). It is ironic and depressing that, as sociolinguists and also most teachers well know, this aim is far more likely to be achieved from a starting point of the study of linguistic diversity and the nature of spoken English.

tain about how to teach a course that is so different from their previous experience' (Bleiman and Webster 2006: 29). A further serious problem at the time when we began our project was a lack of classroom materials on spoken English. This meant that hard-pressed teachers had to make time to locate suitable audio or video clips to use in class to prepare their students for the examinations. Once found, transcribing spoken or visual material for classroom use is equally time-consuming, and not easy for teachers with little or no previous experience. The success of the GCE A-level English Language examination was said in 2006 to rely on a small energetic group of teachers who found their own materials on the Internet or who drew on the media and were happy to share their resources with others (Bleiman and Webster 2006: 29). For many teachers, though, this degree of commitment was likely to be daunting and too time-consuming.

Despite the difficulties of teaching GCE English Language, there are good social reasons why it is worth encouraging its take-up in schools. The subject is especially popular in schools in multicultural urban areas, where the focus on sociolinguistic variation and language and identity appeals to students from minority ethnic backgrounds since their own linguistic experiences are relevant, valued and analysed during their studies. It positions speakers of nonstandard varieties to acquire Standard English more effectively through learning about the social diversity of English (Wolfram 1998: 182). The subject allows less able and average students to achieve as well as the most gifted (Bleiman and Webster 2006: 29), so it provides opportunities for social mobility and for improving the skills base of UK society, with significant benefits for both individuals and society. Furthermore, the subject is 'boy-friendly': 6 per cent of all male and 8 per cent of all female A-level entries are for English Language, compared to 14 and 26 per cent for English Literature (Vidal Rodeiro 2006: Table 3). Boys prefer GCE A-level English Language to other language subjects, so it gives them the opportunity to catch up with girls in the acquisition of language skills.

It was against this background that we decided to produce resources for teachers to use in the classroom based on our recent research on the English of young people in London. This was one of several initiatives that were developed at the time in response to the situation just

described. Others include British Telecom's *All Talk: English 14–19*; the British Library *Sounds Familiar* web pages; *The Talk of the Toon* (Corrigan et al. 2012) web resource, described elsewhere in this volume, and more.

2 The London Projects

The research project *Linguistic Innovators: The English of Adolescents in London* (Kerswill et al. 2004–2007) was the first large-scale sociolinguistic study of English in London, testing the view that long-standing migration patterns make London the origin of language changes in the UK and beyond. It resulted in a transcribed corpus from audio recordings of 1.4 million words from indigenous Londoners aged 70+ and adolescents aged 16–19 from many different ethnic groups. We found that, contrary to thinking at the time, London is actually not the source of the language changes underway in many UK urban centres. Instead, young people in the multicultural inner-city area used a repertoire of innovative features, in all components of language. We refer to this way of speaking as Multicultural London English (MLE) and argue that MLE has replaced 'Cockney' (see also Fox 2015). Young people from immigrant backgrounds led in the use of MLE features, but white speakers from long-standing 'Cockney' families also used them.

Our second London project, *Multicultural London English* (Kerswill et al. 2007–2010), aimed to establish how MLE arose. A further 2 million words were recorded and transcribed from speakers aged 4, 8, 12, 16–19, 25 and 40, including parents and caregivers of 12 of the younger children. We found that MLE was well established among the youngest children, suggesting that they acquired it from peers and siblings, not their parents (who were mostly non-native speakers of English). We concluded that children in multilingual areas of London acquire combinations of language features from a rich 'feature pool' of linguistic forms influenced by a wide variety of languages, dialects and learner varieties. The pool serves as a resource and a model for non-native speakers acquiring English where there is no consistent target variety. This is a new dynamic of change affecting a metropolis containing a large minority ethnic and/or immigrant population, with strong implications for

our understandings of processes of language change; see, for example, Cheshire et al. (2011a).

Although the research had mainly theoretical objectives, there were also some educational aims. As part of the project work, Susan Fox participated in three Knowledge Exchange workshops with teachers and students of GCE A-level English Language, where sound clips from the project recordings were discussed. Teachers were keen to use these recordings in their classes, because of the difficulties mentioned earlier of obtaining appropriate spoken English material. We learned that they were likewise interested in keeping abreast of relevant current research in linguistics, but that this was not easy because of a lack of time and the difficulties of accessing journals. This encouraged us to successfully seek further funding to develop follow-on resources from the two London projects (Cheshire et al. 2011b).

3 The English Language Teaching Resources Archive

We began the follow-on work by setting up an advisory panel of 14 teachers from the north and south of England. Two of the teachers were also experienced examiners for GCE A-level English Language; one of the two also wrote textbooks for both GCE A-level and GCSE English Language courses. The panel advised us at every stage of our work to ensure maximal relevance to the needs and interest of classroom teachers; they also piloted the resources as they were developed. With their guidance, we produced a website containing three types of resource: a Databank of sound clips and accompanying transcripts; a Linguistics Research Digest; and a set of Language Investigations, all interlinked as shown in Fig. 10.1 and as we explain further below.

3.1 A Databank of Spoken English

The Databank contains 12 folders, each with a sound clip taken from the recordings made for the two London projects, a written transcript of the clip, and a set of discussion points. The speakers are from a range

From Sociolinguistic Research to English Language Teaching

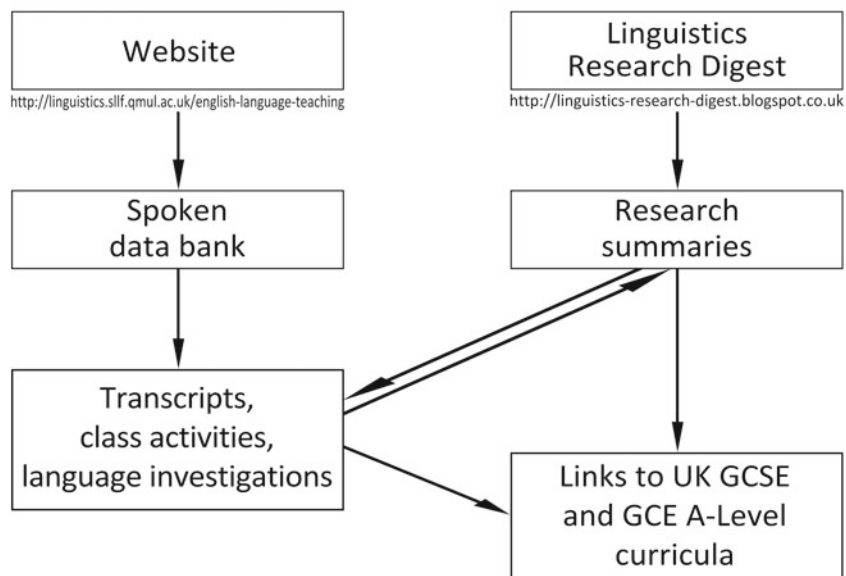


Fig. 10.1 The English language teaching resources archive

of ethnic backgrounds, both male and female, and include 8-year-olds, adolescents and a 77-year-old man. Table 10.1 details the titles we give to the extracts, the speakers (using pseudonyms) and their ages and backgrounds.

The design of the Databank follows the recommendations of Reaser and Adger (2007), which are based on many years' experience of working with teachers. Reaser and Adger point out that materials with descriptions of specific language features are useful resources, and that since teachers are the experts in what is required for the classroom, it is important to provide background information and suggestions so that they can decide for themselves which parts of the materials to use. They caution that teaching materials produced by linguists must be consistent with the goals of the curriculum. We therefore chose sound clips that illustrate specific aspects of spoken English, such as interruption and overlap (Competing for the floor), quotative expressions

Table 10.1 Folders in the Databank of spoken English

Title of extract	Details of speakers
Food stories	Dafne (female, age 8, Nigerian family) and Nandita (age 8, Bangladeshi family)
Competing stories	Derya (female, age 8, Turkish family) and Kareen (female, age 8, Indian family)
The dog story	Howard (male, age 8, white British) and Junior (age 8, Afro-Caribbean)
Competing for the floor	Lydia and Louise (both are female, age 8, white British)
Dressing up	Madeleine (female, age 8, mixed race white British/Afro-Caribbean)
At the airport	Alex (male, age 16, mixed race white British/Afro-Caribbean)
Street trouble	Angela (female, age 16, mixed race white British/Afro-Caribbean)
How Courtney met her boyfriend	Courtney (female, age 17, and Aimee, age 19, both from Jamaican families)
Walking home from cadets	Tina (female, age 18, mixed race white British/Indian)
Problem at college	Laura (female, age 19, white British)
The bike incident	Zack (male, age 16, white British)
Life in the army	Stan (male, age 77, white British)

(The bike incident) or filled and unfilled pauses (Life in the army, and several of the other clips). Some extracts illustrate MLE, as a social and geographic variety of English. The discussion points for each transcript and sound clip draw attention to relevant linguistic features in the clip; and they include most of the linguistic features mentioned in GCE English Language textbooks. We use the linguistic terminology recommended by the Department for Education in the specifications for the English Language GCE examination, plus some additional conventional terminology when needed, and there is a separate glossary in the Databank explaining all the terms.

Example (1) gives, as an illustration of our approach, the introduction to the sound clip and transcript from the Life in the army folder. Examples (2) and (3) show extracts from the discussion points for this folder; line numbers in (2) and (3) refer to the numbered lines in the download version of the transcript.

1. Stan is 77 years old and lives in Havering, Essex. In this extract he reminisces about his army days. It could be interesting to consider which features of Stan's speech mark him out as a member of an older generation: for example, would a young person use the colloquial words and expressions *chap* (lines 6, 75), *a good half hour* (line 25), *a blind bit of notice* (line 31), *blimey* (line 44) or *a great big fat corporal* (line 18)? Stan's hedges, discourse markers and quotative expressions are also more typical of an older speaker. Much of the impact of the story comes from what Stan said to his superior officers and what they said to him, so there is a lot of reported direct speech and reported thought in his story.

2. *er* and *erm* (filled pauses)

These nearly always occur at the beginning of a clause, indicating that Stan wants to keep the floor while planning the grammatical structure of what he is about to say (lines 2, 10, 11, 18, 25, 29, 77). Sometimes there is a silent pause too (lines 11, 25, 29). Stan mainly says *er* (*erm* occurs only once), in line with attested gender differences in the use of *er* and *erm*.

3. Hedges

Sort of (lines 21, before the verb *scratched away* and 30, before the verb *walk about*) involves the listener by signalling imprecision—'scratched away' may not be the best way to describe the sound of an old gramophone, and the impression Stan gives of the way he walked about (line 30) may not be exactly right. *About* (line 75) signals that 30 is an approximate number. Note that although young people use *sort of* and *about* too, in these contexts they may be more likely to use *like*.

3.2 Ethical Considerations

The importance of using research to benefit society has long been recognized by sociolinguists and has been especially pioneered by the two leading US sociolinguists William Labov and Walt Wolfram. Labov's (1982) 'principle of debt incurred' and Wolfram's (1993) 'principle of linguistic gratuity'

both promote the notion that linguists should endeavour to make their research and information about language both available and accessible to the general public as described in some form or other in all contributions to this volume.

The work we describe in this chapter is in line with these principles. It is clearly worthwhile to make corpora such as those compiled from the London projects available for other researchers to use and, equally, to adapt them and make relevant information available to a wider general public. However, there are many ethical issues to take into consideration before allowing access to the data.

First of all, let us consider access to the transcripts. When collecting data during any sociolinguistic study, the participants are guaranteed anonymity and confidentiality of subject matter discussed during the interview. In order to make the transcripts accessible to other researchers or if we want to use examples from the transcripts when explaining concepts to non-linguists, they must therefore be anonymized in order to protect the identity of the speaker, but the question is how far do we need to go in order to ensure anonymity? At the community level, for example, where do we draw the line? In our data, we left in such references as 'Hackney' and 'Havering' (the two London boroughs where the research was conducted). We felt that it was necessary to keep these references in the transcripts because our results cannot be generalized to other areas of London where, perhaps, the demographics differ. The names of London areas such as 'Romford', 'Islington' and 'Wood Green' were also left in the transcripts, as were names of streets if they were used in a general sense as in the examples *I buy my jeans in Mare Street* or *I used to work in a bar down near Liverpool Street*.

Many young people in London refer to the area that they live in by their postcode as in the example *I'm from E8* and these were generally left in the transcripts unless they were mentioned with more specific addresses. However, we anonymized the names of streets when they were used in a more specific sense as in *I live in (name of street)*. We also anonymized names of places where they could be used to track down an individual as in the example *if you play football with us yeh over (name of park)*. We also left out names of schools if a speaker said the name of the school that they had attended or where a general reference to a school could lead to the

identity of an individual as in the example *some white girl from your area ... she goes (name of school) she knows (name of girl)*.

As far as individuals were concerned, we anonymized the names of the speakers and the names of any individuals that were mentioned—teachers' names for instance. Any obvious private information (such as telephone numbers, addresses or names of specific clubs attended) was also removed. However, is this enough to maintain the level of confidentiality and anonymity expected of ethical sociolinguistic researchers? Other issues of confidentiality might revolve around such things as sexual orientation or even sexual activity, highly personal and intimate topics that occur fairly frequently in our data. Should this kind of 'personal' information be removed? In our data (which is currently only available to other researchers) we retained this kind of information, provided we were confident that the speaker was sufficiently anonymous. Similarly, dates of birth were retained, although such decisions would perhaps be different if the transcripts were to be made available to a wider public.

Then there is the consideration of other, more 'public' individuals such as celebrities and TV personalities—is it ethical to leave these in the transcripts or do researchers have a responsibility to also protect public figures if disparaging comments or allegations are made about them? The same applies to more 'local' celebrities, for example the names of locally known music artists (again, particularly when derogatory remarks are made about them). Where do we draw the line?

In a recent volume of *Corpus Linguistics and Linguistic Theory*, Childs et al. (2011) discuss the view that you cannot fully separate the person from the conversation because an interview is inherently personal and people are always invoking their personal experiences and personal opinions even if nothing especially 'personal' (or topics that might be considered more private) comes up. With this in mind, is it therefore even possible to achieve complete anonymity in the transcripts? Even if we take the precautions already discussed above (deleting names, private information and so on) we still find examples such as the following:

4. Interviewer: where did you go to school?

Grant: oh er (name of school). in Archway

Interviewer: is that a Catholic school?

Grant: yeh

As we can see in this example, the name of the school has been deleted. However, the fact that the general area of 'Archway' has been left in and the fact that the school is 'Catholic' actually narrows this down to one easily traceable school. Coupled with other more personal information in the transcript about the speaker's age, ethnic background, details about parents and siblings, it becomes possible that the speaker could be traced and (in a worst-case scenario) some type of harm could come to the speaker as a result of taking part in the study. In our London data we often get references to local gangs, both in the sense of in-group and out-group membership. If they have been referred to in a general sense in the two London corpora then they have, on the whole, been left in but we might see a potential for harm if the corpus is more widely available (to journalists, for example).

Then we find more general descriptions such as *that Chinese up the road down erm near the job centre on Mare Street*, referring to a specific Chinese restaurant. Should these references also be removed or does it depend on the context in which the reference occurs? Questions of this nature then feed into whether it is possible to have an 'objective' anonymization protocol or whether these decisions are always going to be subjective. It is clear that different researchers and transcribers will respond differently to these issues depending on the type of community they are researching and the content of the subject matter contained within their recordings. Matsumoto (2015), for example, in her work on Palauan English, has highlighted how different cultures might consider certain issues as being sensitive, of which the researcher or transcribers might be unaware. Quoting one of her participants, Matsumoto (2015) reports:

When I went to University X in the US and found out how my relatives were quoted in theses, I was really in shock. You know, there're things I swear by God they would never say openly if they'd known their words would be published with their own names. You know, Americans would've thought that we'd never read their theses.

While this quote clearly highlights the need for anonymity, Matsumoto (2015) also notes that researchers need to be aware of 'the existence of locally established beliefs and taboos'. The difficulty that arises here is that the people transcribing the data are often not the researchers them-

selves and these kinds of sensitive issues can easily be overlooked during the transcribing process. Even with the guarantee of anonymity, one might question whether participants would be willing to discuss certain sensitive topics if they were aware that their words might appear in print somewhere in the public domain.

This leads to another important issue: the extent to which the consent form completed by the participants in the projects covers the uses to which the transcripts/recordings will be put. Included here is the situation with regard to children. Children under the age of 16 take part in our projects, with parental consent. However, is this enough? Do researchers then have the right to make the transcripts/recordings of these children available to others? In the consent form for the *Linguistic Innovators* project we state that the recordings will be used for teaching and 'research' purposes only. The issue here is what is meant by 'research'? We feel sure that most of our participants would accept the use of their recordings and their respective transcripts within the more narrow field of scholarly academic research, but can we be sure that they would be equally accepting of 'research' in a wider sense, such as the type of research carried out by other educationists or journalists and the media in general if the transcripts/recordings became more widely available? Furthermore, if the materials become available to the general public, can this still be considered 'research'? In our second project, *Multicultural London English*, the consent form was extended to include permission to use extracts in 'broadcasting' (mainly in response to the many requests from the media to provide sound clips which we were not able to do for the *Linguistic Innovators* project). We also made it clear to the participants that the anonymized transcripts would be kept in an archive for other researchers to use.

Thus far, our discussion has been restricted to the use of the transcripts, but what of the audio recordings? Currently, these are only available to the current research team and to a few members of the academic community who are working with one or more members of the team. In general, the audio recordings present more ethical concerns than the transcripts.

Firstly, we are concerned that the content of the recording, matched with the voice of the speaker, may potentially lead to that speaker's identification, thereby breaching the confidentiality and anonymity guaranteed to the participant. Many of our participants spoke about sensitive topics

on the understanding that the recording would not be heard by anyone outside of the research team. We have many instances of questions and comments such as *This is not going to be played to anyone, is it? Are you sure no-one will hear this?* or of one friend saying to another during the interview *don't worry, it's confidential* when a speaker might be hesitant about discussing a particularly sensitive or taboo topic or if the nature of the discussion involved gossip about another person. Do we withhold such recordings? Do we assume that others would accept the recordings becoming available just because they *do not* make such explicit remarks?

Our second concern relates to the content of the recordings, some of which is highly confidential, sensitive and, in some cases, incriminating (in fact, some of the recordings in the London projects were withheld by the fieldworker and not used for analysis at all because of the sensitive nature of the content). Again, there is the question of the recognition of the speaker's voice and the extent to which the data can be 'cleaned up'.

Anonymizing the audio files is labour-intensive and therefore time-consuming and costly. Given the time frame and financial constraints of many sociolinguistic projects this is not generally possible for large data sets. The question then arises of how this exercise would be funded and, perhaps more importantly, *who* would carry out this exercise and how much knowledge anonymizers would have about relevant local issues. Assuming that a number of different anonymizers would be working on a data set there is the potential for different levels of anonymity to be applied according to individual ideologies—some may have more relaxed views than others. While we would stress the need for an anonymization protocol there will always be grey areas subject to individual decisions.

Finally, once the corpus is in the public domain it then becomes subject to the ethical decisions of other researchers. We would expect the same stringent ethical considerations to be applied among all academic researchers, but there is still the concern that researchers may not be aware of relevant local issues and we almost certainly cannot guarantee the actions of a wider 'researching' public. Even within the academic community we have already found that researchers in other countries have not always been trained in research ethics and do not feel the same ethical responsibilities towards participants.

Having highlighted some of the problems that we consider to be of importance in the sharing of data, we nevertheless acknowledge that the benefits of sharing data among researchers and disseminating language information to a wider general public are extremely worthwhile. So far, we have deposited the *Linguistic Innovators* transcripts with the UK Data Service. The corpus (available as concordance-searchable text files) is available to researchers, teachers and students from any field, organization or country on registration with the UK Data Service. Some data sets have restricted access including, as yet, the MLE transcripts, which are available to other researchers only on request to a member of the research team.

Given the concerns raised in this chapter about the audio files, the sharing of this data remains at the discretion of the research team. We are still considering whether there is any way of making the audio files more widely available but have, so far, rejected the possibility of archiving the recordings with an organization such as the UK Data Service.

4 The Linguistics Research Digest

We designed the Linguistics Research Digest as a way of meeting teachers' desires to keep abreast of research in linguistics. We chose articles published in recent linguistics journals on topics relevant to the specifications for GCE A-level and GCSE English Language, and summarized them in a way that aimed to be engaging and jargon-free as well as accurate. The teacher advisory group advised us on the choice of journal articles, the style of the summaries (in terms of their accessibility, interest and their form), and the frequency with which summaries should be posted. We posted the summaries on a blog site—two summaries a week during the life of the Follow-on project, in 2011, and one summary each week from 2012. Fig. 10.2 is a screenshot illustrating the Digest.

At the end of the summary we provide the full bibliographic reference to the article, and there are hyperlinks in the summary itself to the web pages of the authors. As Fig. 10.2 shows, the side bars contain links to other language blogs and to sites relevant to the English Language curriculum, and a searchable tool for browsing the Digest by category.



Linguistics Research Digest

Blogging on language issues

Home About Contact

Follow The Linguistics Research Digest by Email

Email



MA in Linguistics at Queen Mary

Browse digest by Category

- Accents and Dialects
- Bilingualism
- Child Language
- Cognition

Thursday, 12 January 2012

Uh, more on the mysterious case of 'uh' and 'um'



A recent summary on this blog (*Er, what about this?*) discussed the intriguing finding that while male speakers of British English used *um* and *uh* (or *erm* and *er*, in British English) more often than female speakers, females preferred *um* over *uh*. Now recent research in the US has revealed that female speakers of American English behave in the same way – at least in the two sets of data that Eric K Acton analysed.

Awards

- BAAL Applying Linguistics Fund 2012

Sponsors of the Linguistics Research Digest

LAGB

LAGB - the leading professional association for academic linguists in Great Britain



The British Association for Applied Linguistics

Related Links

- English Language Teaching

Fig. 10.2 The Linguistics Research Digest: first part of a summary

The transcripts and discussion points in the Databank contain links to summaries that relate to specific linguistic features. For example, the discussion points for the Life in the army folder include Stan's use of filled pauses [see example (1) above] and are linked to the Digest summary illustrated in Fig. 10.2. Fig. 10.3 shows how this is done in the Databank folder.

5 Language Investigations

We had intended to provide activity sheets for use in the classroom with the Databank, but the teacher advisory panel found that the sound clips and discussion points provided enough material for their lessons. They

The screenshot shows the website for the Department of Linguistics at Queen Mary University of London. The page title is "Stan: Life in the army". The left sidebar contains a "Main menu" with categories like Home, People, Undergraduate, Postgraduate, Research, Current grants, and Recent grants. The "Recent grants" section is expanded to show "Sociolinguistics Research Group" and "English Language Teaching Project". Under "English Language Teaching Project", "Language investigations" and "Language materials" are listed. The main content area includes a "Sound clip" section, a "Download" link for "Life in the army.WAV", "Discussion points", "Digest links" with two URLs, and a detailed analysis of the text "Stan" under the heading "Clause combining". The analysis notes that "And" is the most frequent conjunction, occurring in lines 4, 5, 11, 14, 19, 20, 21, 22, 23, 25, 27, 29, 33, 35, 37, 40, 41, 45, 47, 52, 53, 55, 56, 57, 65, 67, 68, 69, 70, 77, 82, 84, 86, 87, 90, 92, 93, 98 and 99. It also discusses "Conversational Historical Present" and "Discourse markers" like "oh" and "well".

Fig. 10.3 First page of one of the Databank documents

suggested that instead we developed some Language Investigation tasks. In 2011, when the website was set up, students were required to carry out a small-scale piece of research themselves—a Language Investigation. Although for A-level examinations from May/June 2017 the methods of assessment will change to largely formal written examinations, 20 per cent of the final grade will still be obtained from non-examination based assessment of ‘Language in Action’. This part of the curriculum aims ‘to allow students to explore and analyse language independently and develop and reflect upon their own writing expertise’ (AQA AS and A-level specifications 2014: 18). The latter aim requires students to produce a 1500 word piece of original writing and commentary; the former requires them to produce a 2000 word report (excluding data) of a Language Investigation that they have carried out themselves.

So far we have produced seven Language Investigations, all linked both to and from specific summaries in the Linguistics Research Digest. The investigations were piloted by the advisory group and then revised to

take account of feedback from teachers and their students. The Language Investigations give clear directions on data collection and suggestions for how students can analyse the data. The topics are: giving place directions; speech style in call centres; language brokering; compliments; intensifiers; general extenders; and second person plural forms. As an illustration, Fig. 10.4 shows the Language Investigation on the latter.

6 Use of the Resources

Between January 2013 and 31 July 2013 the Archive received 18,000–20,000 visits per month, of which 60 per cent were return visits; in the same period the Digest received 7000–8000 hits per month. Google Analytics showed that approximately 60 per cent of the visitors to the Archive were from the UK, but that others originated in more than 30 different countries, from all continents. Since its inception, the Digest has attracted over 300,000 hits, with the highest number of visitors coming from the USA, closely followed by the UK and then, in descending order among the top ten countries, France, Germany, Russia, Ukraine, India, Australia, Canada and China. Websites set up by schoolteachers recommend the resources: for example, one teacher writes ‘There is a fantastic blog produced by Queen Mary University of London’s Linguistics Department. They have a real commitment to encouraging A-level students in their study of the English language ... even better, they have come up with some possible A2 level investigations and for some, they even suggest a methodology and research question ... they have even provided access to a whole data bank of spoken contemporary London English ... this is a fantastic opportunity’.³ We posted an online survey in June 2013 asking for feedback on the Resources Archive. The following quote is typical of the feedback we received from the survey: ‘The audio and transcripts have been invaluable in helping me prepare students for the exams and coursework, and the glossary of terms has always been a handy reference point’. We discovered that examiners used the Linguistics

³ See <http://eastnorfolklanguage.blogspot.fr/2013/06/great-leads-for-possible-investigations.html> (accessed 8 August 2015).

Language Investigations in spoken English

"Do you understand who I'm talking to?" Second person plural forms in English



Most languages have separate words for singular and plural pronouns. English used to have separate second person pronouns too, but since *thou* fell out of use the *you* pronoun has had to do double duty. So, in the scene depicted here, is the speaker accusing one of his friends, or all of them? How do we deal with this problem?

you took my biscuit!

You could investigate how English speakers make clear who they're talking to when there is more than one person around. How do they show that they are speaking to just one person? Or to two people? Or to the whole group?



How to investigate?

Listen and note

One way to find out is to listen to what your teachers say when they are addressing one person, the whole class, or a small group of students. Note this down during the course of a day, so that you end up with a collection of phrases. Perhaps the teacher uses the student's name as well as *you* (but this obviously wouldn't be possible for a group

Fig. 10.4 Example of a Language Investigation

of students!) You'll probably find that the phrases include *you all* (for example, *will you all now think about this?*) and *the two of you* or *both of you* (for example, *would the two of you do this?*) What other phrases does your teacher use to show who he or she is addressing? Does the teacher sometimes make it clear through eye contact, or pointing?

Some linguists* have claimed that we have an unconscious rule about how to address two or more people: if it isn't clear from the context, the speaker must make it perfectly clear whether they are referring to one person, to everyone who is there or to a subset of the people who are there. They usually do this by using people's names, or a phrase like *all of you*, *you fellows*, *both of you*, or by gesturing (usually pointing). Once this has been made clear, it is OK to use *you* from then on, but only until the next ambiguous moment in the conversation. If someone joins the group or if someone leaves, the speaker has to make it clear all over again just who they are talking to.

[* Andrew Pawley and Frances Hodgetts Syder (1983) Natural selection in syntax: notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics* 7: 551-579.]

Watch TV

Researchers have found that in the *Friends* series, the speakers often use *you guys* when they are addressing more than one person. You could watch an episode of *Friends* and note down all the words and phrases used when people address more than one person. How often do speakers say *you guys*? Are there any other words or phrases that they use to make it clear whether they are talking to one person or more than one person? Is it always clear what *you* means?

Perhaps more interestingly, watch a British TV sitcom where people sometimes address more than one person (such as *Big Brother*). Do people use *you guys* here? If not, how do they make it clear who they are addressing?

Do some dialect research

Fig. 10.4 (continued)

Many varieties of English have a separate second person plural pronoun, unlike standard English. There are many different forms, including *youse*, *you all*, *yinz* or *you uns* (and more). You could browse the internet or look at some Linguistics textbooks to gather examples. Find as many second plural pronoun forms as you can, and note down in which parts of the world they are heard.

TIP:

Try googling "second person pronouns" and national varieties of



English around the world, such as "Irish English", "America", "Australia", "Jamaica" or "South Africa". You could also search for second person pronouns in regional varieties (dialects) of British and American English.

You could also try to find out what has happened to the old singular pronoun *thou*. Is it still used? If so, where?

In conclusion

Once you've done your investigation, consider whether *you* in English is really ambiguous. Do people really not know who is being addressed when they hear *you*? Or do they find other ways of showing that they are addressing more than one person?

Suggested Reading:

Theresa Heyd (2010) How you guys doin'? Staged orality and emerging plural address in the television series *Friends*. *American Speech* 85 (1): 33-66. (Click [here](#) for a summary of this paper).

Fig. 10.4 (continued)

Research Digest with teachers at training sessions ‘to encourage them to find new material to point students towards and it’s always gone down very well’.

We learned, too, that the resources were useful for teachers and students of EFL/ESL and university-level English Language in the UK and beyond. A typical response from the survey is the following: ‘I teach linguistics, working with prospective teachers—and practicing teachers—in the USA. This is a terrific site for them to look at ... and to talk about the phenomena here in the context of analogous topics in varieties of American English they know and are likely to encounter’.

7 Workshops for English Language Teachers: *Analysing Spoken English*

A further spin-off of our project for producing resources for teachers was the organization, in collaboration with Dr Heike Pichler (Newcastle University), of workshops for teachers of English Language. The workshops, held in April 2012 at the University of Salford, July 2012 at Queen Mary University of London and December 2012 at Newcastle University, aimed to disseminate insights from scholarly research about language variation and change and to provide teachers with an overview of data-banks and resources available online for use in the classroom. These one-day workshops consisted of talks from the organizers and other invited researchers and consisted of two parts.

In the first part of the workshop we aimed to break down persisting prejudices against the use and users of discourse-pragmatic features such as *innit* (as in *It’s only an hour from Edinburgh and Newcastle, innit?—Oh, I’ve answered this one before, innit?*), quotative forms such as *be like* (for example, *And they were like, ‘we divn’t want you here’. And we were like, ‘why?’*) or intensifiers such as *dead* (as in *It was dead funny*). The aim was to demonstrate that it is wrong to dismiss discourse-pragmatic features as mere fillers which contribute nothing to the content or communicative force of an utterance or, even worse, to perceive them as a sign of inarticulateness, laziness or lack of intelligence. We aimed to break down

existing prejudices against the use of such features and to demonstrate how these features develop, what communicative function they perform in interaction (such as to signal tentativeness or assertiveness, or to facilitate speaker change), and how they change over time. We thereby hoped to raise participants' awareness of discourse-pragmatic features and to demonstrate that they play a vital role in interaction.

In the second part of the workshop, we provided teachers with an overview of currently available resources for working with spoken data in the classroom, focusing in particular on a demonstration of two projects specifically aimed at providing teachers with relevant resources. The first, the *Diachronic Electronic Corpus of Tyneside English* (DECTE) is a corpus of spoken language from north-eastern England spanning five decades. The linked *Talk of the Toon* website developed by the same project team which is more fully described elsewhere in this volume is a multi-media publicly-available resource containing audio recordings and transcriptions as well as still and moving images relating to themes relevant to subject areas in the National Curriculum. The second project is the one discussed in this chapter, *The English Language Teaching Resources Archive*, which focuses on London English and had the aim of developing accessible classroom materials arising from sociolinguistic research on spoken language.

We also compiled a Resource Booklet for the teachers, which contained factsheets summarizing relevant insights from current research on the selected topics as well as relevant scholarly articles that teachers would find useful. The materials included suggestions for classroom activities that would enhance students' theoretical knowledge about spoken language and language variation and change. We provided answers and commentaries to these activities. We also included photocopyable worksheets for practical investigations into spoken language, similar in nature to the Language Investigations described earlier.

The workshops were very well attended, with around 30 participants at each one. At the end of each workshop we asked participants to complete an evaluation questionnaire. The value of such workshops for teachers is captured in some of the comments they made on the questionnaires, examples of which are given below:

5. The presentations had a very clear sense of the needs and level of understanding of the audience and was [*sic*] consequently very accessible and engaging. Also—a very coherent focus to the whole event.
6. Very interesting to hear about recent research. Made relevant to our A-level teaching context. Excellent resource pack.
7. Up-to-date information/details about recent research. The website resources look really useful especially the recordings/transcripts/condensed research data.

Overwhelmingly, many of the teachers asked for further sessions to be organized on the same and other topics. They asked for the workshops to be provided in different locations around the UK and there were many requests for similar workshops to be made available to students of A-level English.

8 The Future

As we mentioned earlier, the English Language Teaching Resources Archive was developed with the help of external research funding for one year. Further funding is now needed to develop the resources further. For example, we would like to extend the Databank so that it contains sound clips and accompanying transcripts from other regional or indeed national varieties of English, and we would like to post additional language investigations based on articles summarized in the Research Digest. Even without this, however, the existing Databank remains as an online resource, and it continues to be used.

Of course, lack of funding not only hinders development of the resources; it could also mean that it is difficult to sustain the website and to deal with any technological problems that may arise. However, UK higher education institutions are subject to a national assessment of the quality of the research carried out by their academic staff, the results of which determine part of the government funding given to each institution. The most recent assessment system, the Research Excellence Framework (REF),

includes assessment of the impact of research.⁴ As a result, many universities have now set aside a budget to encourage public engagement with the research of their staff, and to ensure that relevant ‘users’ are able to access the research. The Databank of spoken English has benefited from this as the university where the resources were developed (Queen Mary University of London) was willing to provide the financial resources and the person power for the website to be made part of the web page of the Department of Linguistics. This means that it can be maintained along with the department’s web pages, and its future is assured.

We have been fortunate in obtaining financial sponsorship for the Linguistics Research Digest from the Linguistics Association for Great Britain and the British Association for Applied Linguistics, so far every year since 2012; and in 2012 the Archive won further financial support from the British Association for Applied Linguistics’ ‘Applying Linguistics’ competition. We use the funds to recruit able graduate students, who try their hand at writing summaries of relevant journal articles and thereby gain experience of writing for a lay audience. We edit these summaries and also write some ourselves but, as is the case for the teachers for whom the Digest was developed, time and resources are in short supply. Our digest was modelled on the British Psychological Society’s very successful Research Digest,⁵ which is maintained with the support of a permanent part-time post. We have not so far been able to secure a post of this kind for work on the Digest, but we continue to explore possibilities.

References

Books and Articles

- AQA Education. 2014. *AQA AS and A-level English Language Specifications*. Version 1.0, 14 October 2014. Manchester: AQA Education.
- Bleiman, Barbara, and Lucy Webster. 2006. *English at A level: A Guide for Lecturers in Higher Education. A Report to the Higher Education Academy*

⁴ For further details, see the REF website at <http://www.ref.ac.uk>.

⁵ See <http://digest.bps.org.uk>.

- English Subject Centre*. London: English and Media Centre Report Series no. 12.
- Cheshire, Jenny, Susan Fox, Paul Kerswill, and Eivind Torgersen. 2011a. Contact, the feature pool and the speech community: The emergence of Multicultural London English. *Journal of Sociolinguistics* 15(2): 151–196.
- Cheshire, Jenny, Susan Fox, and Paul Kerswill. Jan–Dec 2011b. *From Sociolinguistic Research to English Language Teaching*. UK Economic and Social Research Council project RES 189-25-0181.
- Childs, Becky, Gerard Van Herk, and Jennifer Thorburn. 2011. Safe harbour: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory* 7(1): 163–180.
- Department for Education. 2013. *English Language GCSE Subject Content and Assessment Objectives*. London: Department for Education DFE-002320-2013. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254497/GCSE_English_language.pdf (accessed 9 June 2015).
- . 2014. *GCE AS and A level Subject Content for English Language*. London: Department for Education, DFE-00362-2014. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/302109/A_level_English_language_subject_content.pdf (accessed 9 June 2015).
- Fox, Susan. 2015. *The New Cockney: New Ethnicities and Adolescent Speech in the Traditional East End of London*. Basingstoke: Palgrave Macmillan.
- Kerswill, Paul, Jenny Cheshire, Susan Fox, and Eivind Torgersen. 2004–2007. *Linguistic Innovators: The English of Adolescents in London*. UK Economic and Social Research Council project RES 000 23 0680.
- . 2007–2010. *Multicultural London English*. UK Economic and Social Research Council project RES 062 23 0814.
- Labov, William. 1982. Objectivity and commitment in linguistic science. *Language in Society* 11(2): 165–201.
- Matsumoto, Kazuko. 2015. Beware that the community changes: Issues of the returns for linguistic favours. Unpublished manuscript. University of Tokyo.
- Reaser, Jeffrey, and Caroline Temple Adger. 2007. Developing language awareness materials for non-linguists: Lessons learned from the *Do You Speak American?* Project. *Language and Linguistic Compass* 1(3): 155–167.
- Vidal Rodeiro, and Carmen L. 2006. *Uptake of GCE A-level Subjects in England 2001–2005*. Statistics Report Series no. 3. Cambridge: Cambridge Assessment, Local Examinations Syndicate, University of Cambridge.

Available at <http://www.cambridgeassessment.org.uk/Images/109897-uptake-of-gce-a-level-subjects-in-england-2001-2005.pdf> (accessed 8 August 2015).

- Wolfram, Walt. 1993. Ethical considerations in language awareness programs. *Issues in Applied Linguistics* 4(2): 225–255.
- . 1998. Dialect awareness and the study of language. In *Students as Researchers of Culture and Language*, eds. Ann Egan-Robertson, and David Bloome, 167–190. Cresskill, NJ: Hampton Press.

Websites and Online Resources

- British Library. *Sounds Familiar*. <http://www.bl.uk/learning/langlit/sounds>. (accessed 8 August 2015).
- British Telecom. *All Talk: English 14–19*. <http://www.btplc.com/BetterFuture/ConnectedSociety/LearningAndSkillsFreeResources/AllTalk> (accessed 8 August 2015).
- Cheshire, Jenny, and Susan Fox. 2011a. *English Language Teaching Resources Archive*. <http://linguistics.sllf.qmul.ac.uk/english-language-teaching> (accessed 8 August 2015).
- . 2011b. *Linguistics Research Digest*. <http://linguistics-research-digest.blogspot.co.uk> (accessed 8 August 2015).
- Corrigan, Karen P., Isabelle Buchstaller, Adam Mearns and Hermann Moisl. 2012. *The Talk of the Toon*. Newcastle University. <http://research.ncl.ac.uk/decte/toon> (accessed 8 August 2015).
- REF. 2014. Research Excellence Framework. <http://www.ref.ac.uk> (accessed 8 August 2015).
- The British Psychological Society Research Digest*. <http://digest.bps.org.uk> (accessed 8 August 2015).

11

Analysing Spoken Discourse in University Small Group Teaching

Steve Walsh and Dawn Knight

1 Introduction

Our main argument in this chapter is that there is a strong rationale for combining corpus linguistics (CL) with conversation analysis (CA) in many settings, but especially in educational contexts. The main reason for adopting this position is that CL is unable to account for some of the features of spoken interaction which occur at the levels of utterance and turn, while CA is unable to handle larger data sets. Although CA and CL have both been used independently to study spoken encounters, each has its limitations. For example, in many situations, CL largely ignores context and focuses on large-scale analysis, whereas CA offers detailed, microanalytic descriptions but is unable to generalize to larger contexts. Using a combined CL and CA approach (henceforth, CLCA), we argue, cumulatively gives a more ‘up-close’ description of spoken interactions

S. Walsh
Newcastle University, Newcastle upon Tyne, UK

D. Knight (✉)
Cardiff University, Cardiff, Wales

than that offered by using either one on its own. From the analysis, we can gain powerful insights into the ways in which interactants establish understandings and co-construct meaning through the use of particular words, utterances and social actions.

For the purposes of this chapter, we talk about CL as a *methodological tool* which will help us investigate a corpus of small group interactions recorded in higher education (HE). Using CL as a tool allows us to automatically search a large data set, something which would have been impractical manually. However, while CL allows us to count frequencies and find keywords in microseconds, thus revealing patterns that we could not otherwise find, it does not allow us to explain the dynamics of these interactions. CL does not, for example, tell us how meanings are established, how repair is managed and how conversational trouble is dealt with. For this, we need a much finer-grained approach to analysis, such as that offered by CA.

Increasingly, CL is being applied to contexts and domains outside of the study of language itself where the focus is on the *use* of language in a given context. Such contexts include courtrooms and forensic linguistics (Cotterill 2010), the workplace (Koester 2006), educational contexts (O’Keeffe and Farr 2003; Walsh and O’Keeffe 2007), political discourse (Ädel 2010) and the media (O’Keeffe 2006), among other areas. Very often, then, CL is used as a tool and another approach, such as CA, critical discourse analysis (CDA) or pragmatics, is drawn on as a framework. To call CL a methodological tool is not to denigrate it. None of the above studies could have achieved the same insights without CL. Essentially, we are distinguishing between *pure* and *applied* CL research. Pure CL research has as its main focus a description of the language in the particular corpus under study, while applied CL research looks at the wider interactional context of language in use. In applied CL research, which is what we are concerned with here, the corpus and its description are not an end in itself; the corpus is merely a means to the end of finding out more about a broader research question.

We also note the salience of a combined CLCA methodology for research into spoken discourse. As more and more modestly sized specialized corpora emerge, we see an increasing use of CL with other

approaches to the analysis of discourse in context (see Santamaría-Gracia 2010, for example). As O’Keeffe and McCarthy (2010) point out, in the early days of CL, the aim was to have very large written corpora to serve the needs of lexicographers, whose focus was obviously on semantic and lexical patterning rather than on discourse context. The resulting large corpora were lexically rich but contextually poor. That is to say, lexical items in a mostly written corpus of 100 million words or more are, in the main, detached from their context. In the case of smaller, specialized and highly contextualized corpora (such as the one used in the present study), there is the potential to extend the analysis beyond lexis to areas of use, with a focus on issues relating to pragmatics, interaction and discourse.

In the present study, our first pass at the data enabled us to scope out and quantify recurring linguistic features, using CL. By linking these recurring features in the corpus to the local context, we were led to contextual ‘patterns’—features in the data which seemed particularly prominent. The second layer of analysis (using CA) draws upon these contextual patterns and investigates them more closely by looking at, for example, turn organization and sequence, overlaps, latching and so on. The process was non-linear in that we sometimes used CL tools within the CA layer of analysis to quantify CA insights. We can say that our analysis progresses in an iterative manner: from CL to CA, back to CL and so on. There is an interdependence between the two modes of analysis.

Having established a position on the suitability of a combined CLCA methodology, we now turn to a consideration of the context in which the study took place.

2 Context: Small Group Teaching in HE

In many HE settings, small group teaching (henceforth SGT) contexts such as seminars and tutorials are used to support lectures by allowing tutors and students to engage in discussion and debate. To take the example of one subject, psychology, SGT can account for around 40 per cent of the contact time of first- and second-year undergraduates

and up to 75 per cent of final-year undergraduate and postgraduate students (Bennett et al. 2002). From the perspective of CL, much influential work on spoken interaction in HE is based on corpora such as MICASE (*Michigan Corpus of Academic Spoken English*) (Simpson et al. 2002) or BASE (*British Academic Spoken English*) (Thompson and Nesi 2001). Both corpora comprise data from a range of speech events in HE, including contexts relevant to the study reported here, such as classroom discussions, seminars, lab work and advising sessions. Studies based on the MICASE corpus have explored a wide range of phenomena in academic spoken interaction, such as metadiscourse in lectures (Lorés 2006), and the use of conditionals (Louwerse et al. 2008).

Outside CL, recent research on talk-in-interaction in SGT in HE has uncovered important aspects of the processes or 'machinery' by which seminars and tutorials 'get done'. Such work has focused on cues and signals used to manage interaction and participant roles (Viechnicki 1997), sequential organization and negotiation of meaning (Basturkmen 2002), the issue of 'topicality' in small group discussion (Stokoe 2000; Gibson et al. 2006), and the formulation and uptake of tasks and resistance to 'academic' identities (Benwell and Stokoe 2002). Much of the more recent work on talk in SGT (particularly that of Benwell and Stokoe) draws on perspectives from ethnomethodology, CA and discursive psychology. In these perspectives, SGT sessions are seen as locally produced accomplishments in which participants take actions to further their own goals and agendas, display their orientations to others' actions and make relevant certain identities. In SGT contexts, tutors will demonstrably orient to the accomplishment of pedagogical goals and tasks, and students may accept or resist these actions (Benwell and Stokoe 2002). At all times during interaction in these SGT contexts, as in other educational contexts, there is a complex relationship between pedagogic goals and the talk used to realize them (Seedhouse 2004; Walsh 2011). By looking closely at the interactions taking place, we show that tutors and students engage in tightly organized and intricate negotiations of a set of pedagogic agendas, and in doing so, use as tools both the machinery of interaction (Levinson 2006), such as turn-taking and exchange structure, and specific linguistic features, such as discourse markers, to achieve their goals.

3 Methodology

3.1 *Newcastle University Corpus of Academic Spoken English*

NUCASE (*Newcastle University Corpus of Academic Spoken English*) comprises 1.0 million words (around 120 hours) of academic spoken English recorded across three faculties: Humanities and Social Sciences (HaSS), Science, Agriculture and Engineering (SAgE) and the Faculty of Medical Sciences (Medicine) at Newcastle University. In addition, approximately 25 per cent of the corpus is based on recordings from pre- and in-session English language classes recorded in INTO, the English Language Centre for the university. The data, using video- and audio-recordings taken from seminars, tutorials, PhD supervisions, staff–student consultations, English language classes and sessions involving informal learner talk, are used to provide a ‘snapshot’ of spoken academic discourse in contexts where interaction takes place. Lectures have been excluded for this reason and the main focus is on SGT sessions.

Our rationale for selecting data from each of the three faculties is to consider how orientations to knowledge, learning and skills development vary from one discipline to another. Broadly, we are interested in the ways in which learning ‘gets done’ in different academic disciplines, and how language, interaction and learning combine in that process.

A key aim of the study is to describe, characterize and operationalize interactional competence in an HE setting. We are especially interested in the ways in which tutors and students create meanings and establish mutual understanding in HE spoken discourse. The study will characterize interactional competence, specifically classroom interactional competence (CIC, Walsh 2011) in university SGT and learning sessions. It is anticipated that the wider NUCASE project will focus mainly on CEFR levels B1, B2, C1 and C2 since these are the levels where decisions are made by universities regarding student entry to both undergraduate and postgraduate degree programmes (CEFR: Common European Framework of Reference, a matrix of language descriptors used by language-testing agencies across Europe and beyond to set the standard and

identify descriptors for language testing). These levels correspond approximately to the IELTS (International English Language Testing System) band range 5.5–7.0. We anticipate that the outcomes of the research will have direct relevance to admissions tutors and to language-testing agencies.

Interactional competence has been chosen as an area of focus since it portrays more accurately what goes on *between* speakers rather than a speaker's individual performance. Arguably, the current CEFR descriptors for speaking place more emphasis on individual performance than on interactional competence. It is our intention here to address this imbalance by highlighting the need for descriptors which capture the joint enterprise of any social interaction. In this project, we aim to describe the competence needed to participate effectively in any interaction, rather than on a student's ability to perform, which is often measured under fluency, accuracy, lexical range and so on.

3.2 The Approach

In our approach to analysis, we utilize a CLCA methodology to study academic interactions, as evidenced by data taken from the NUCASE corpus.

The study aims to examine some of the interactional features that emerge as being distinctive of a particular type of talk used in and across a subcorpus of data from NUCASE, which represents data from subjects in each of the three academic faculties represented in NUCASE. These include Bioinformatics (Medicine), Business (HaSS) and Marine Engineering (SAGe). The contents of NUCASE are detailed in Table 11.1.

The process of analysis is the iterative model described in Sect. 1 which combines methods used in CA and CL as a means for offering new insights into the ways in which words, utterances and texts combine to create meaning in SGT in an HE context.

The first parse of the analysis is corpus-driven, involving an examination of some of the most frequently used discursive words and phrases for each academic subject and, where relevant, the identification of

Table 11.1 Contents of the NUCASE subcorpus

Subject	Number of recordings	Hours (approx.)	Class sizes	Approx. word count
Business	79	56:48	4–12 (7.7 average)	628,277
Bioinformatics	24	32:30	4–12 (7.6 average)	143,122
Marine Engineering	19	24	5–13 (6.7 average)	294,839

the most common thematic (semantic) associations of such words and phrases. These basic lines of enquiry set out to isolate emerging patterns in the use of language that exist within and across the subjects. This first layer of analysis provides us with a ‘way-in’ to a subsequent, microanalytic, CA based investigation of the data, which studied the intricacies of these more ‘general’ patterns. From this point, we approached the data in an iterative fashion, working from CL to CA and then back to CL again.

The CL analyses were conducted using Rayson’s (2003) *WMatrix* software. *WMatrix* includes utilities for carrying out word, cluster and parts of speech queries [centring on the production of keyword lists and keyword-in-context (KWIC) outputs]. These analyses enabled us to statistically assess the patterned use of these features in a corpus. With the use of the *WMatrix* semantic tagger (which is based on the UCREL¹ semantic analysis system, see Wilson and Rayson 1993), common themes and semantic associations connected with corpora can also be interrogated using the software.

Descriptive statistics—that is, raw frequency and relative frequency rates (providing a ratio of use based on a ‘per 100 word’ basis, that is, a percentage use)—were used as the basis of the CL analysis, along with log-likelihood (LL) scores. LL scores allow for statistical comparisons of data sets to be carried out, indicating whether differences in keyword frequencies are likely to exist by chance or not. For the purpose of the preliminary CL analysis carried out in the next section of this chapter, only ‘+’ LL scores have been used. These indicate that a particular rate of use is statistically higher in the first subject compared to the

¹ The University Centre for Computer Corpus Research on Language, Lancaster University.

other parameters defined (that is, the other data set to which it is being compared), to a p value of <0.0001 (that is, a critical value of 15.13 or above).

Our initial research questions, informing the preliminary CL analysis, are stated below:

1. Which discursive features emerge as being particularly characteristic of the business, bioinformatics and marine engineering data sets?
2. How do the features identified in (1) compare/contrast from one educational context to another?

3.3 CL Analysis

As a starting point to our analysis, we identified frequency lists of the top 20 words that feature in each subject, as shown in Table 11.2.

Unsurprisingly, function words, rather than content words, proliferate here. Of these function words, we see that the use of personal pronouns is shown to be particularly frequent across all of the subjects (shown in boldface), although the specific rate of use naturally fluctuates from one subject to the next. The first person pronoun *we* is the most frequently used pronoun in the Business data, occurring at a rate of 2.33 times per 100 words, while occurring only 1.31 and 1.69 times per 100 words in the Bioinformatics and Marine Engineering corpora respectively. *I*, in comparison, is used at a rate of 1.45 times per 100 words in the Business data and 2.02 and 2.3 times per 100 words for Bioinformatics and Marine Biology respectively. In addition, *I* ranks higher in the frequency list than *we* for both of these two data sets.

Collectively, as a word class, pronouns are used significantly more frequently in the Bioinformatics data than in the other subjects. The semantic profiling tool in *WMatrix* indicates that pronouns occur in the former data set with a relative frequency of 16.45, 14.52 in the Marine Engineering data set and 14.54 in the Business data. A list of the top 20 most frequent pronouns and their (relative) frequencies across the 3 subjects is shown in Table 11.3.

Table 11.2 The top 20 words in each of the subjects

Rank	Business			Bioinformatics			Marine Engineering		
	Word	Freq.	Rel.	Word	Freq.	Rel.	Word	Freq.	Rel.
1	the	24403	3.88	the	5437	3.8	the	13388	4.54
2	and	15203	2.42	it	3700	2.59	<i>yeah</i>	7510	2.55
3	we	14643	2.33	you	3595	2.51	and	6832	2.32
4	to	12041	1.92	I	2884	2.02	I	6778	2.3
5	you	11610	1.85	<i>yeah</i>	2813	1.97	it	5759	1.95
6	that	11318	1.8	that	2682	1.87	you	5707	1.94
7	it	10453	1.66	and	2631	1.84	to	5410	1.83
8	a	10429	1.66	to	2532	1.77	we	4981	1.69
9	<i>yeah</i>	10242	1.63	a	2133	1.49	a	4853	1.65
10	of	9245	1.47	so	2069	1.45	that	4719	1.6
11	I	9140	1.45	do	1964	1.37	of	4034	1.37
12	so	7711	1.23	we	1877	1.31	is	3819	1.3
13	is	7598	1.21	just	1873	1.31	so	3554	1.21
14	do	6556	1.04	its	1863	1.3	do	3229	1.1
15	in	6473	1.03	is	1760	1.23	its	2723	0.92
16	this	5332	0.85	like	1703	1.19	just	2610	0.89
17	like	4776	0.76	this	1529	1.07	be	2494	0.85
18	what	4723	0.75	what	1421	0.99	this	2356	0.8
19	just	4466	0.71	of	1373	0.96	what	2274	0.77
20	but	4439	0.71	in	1220	0.85	in	2197	0.75

While personal pronouns, including *he, I, you, she, it, we, they*, are, crudely speaking, most commonly used across all of the subjects, the use of the first person singular is more prevalent in the Marine Engineering data (in boldface), while the first person plural is more frequently used in the Business data (italicized boldface, including *we, our, us*). The third person singular (underlined) and third person plural (underlined in boldface) are used in all of the data sets, although the relative frequency with which they are used indicates that they are more frequent in the Bioinformatics data (particularly with the prevalent use of *it*). The use of personal pronouns in general (but first person pronouns specifically) is typically seen as being a characteristic of less formal, spoken, discourse so is an interesting result to find here (see Chafe and Danielewicz 1987; Biber 1992; Biber et al. 1999; Heylighen and Dewaele 2003; Carter and McCarthy 2006; Atkins 2011; Knight et al. 2013, 2014).

Table 11.3 The most frequent pronouns and their (relative) frequencies across each of the subjects

Business			Bioinformatics			Marine Engineering		
Word	Freq.	Rel.	Word	Freq.	Rel.	Word	Freq.	Rel.
<i>We</i>	14643	2.33	<u>It</u>	3683	2.57	I	6768	2.30
You	11608	1.85	You	3595	2.51	<u>It</u>	5734	1.94
<u>It</u>	10390	1.65	I	2884	2.02	You	5707	1.94
I	9132	1.45	<i>We</i>	1877	1.31	<i>We</i>	4971	1.69
That	7235	1.15	<u>Its</u>	1863	1.30	That	3143	1.07
<u>Its</u>	4416	0.70	That	1792	1.25	<u>Its</u>	2723	0.92
What	4291	0.68	What	1318	0.92	What	1952	0.66
They	3955	0.63	This	903	0.63	This	1266	0.43
<i>Our</i>	3388	0.54	<u>He</u>	562	0.39	They	1055	0.36
This	2995	0.48	One	530	0.37	Your	740	0.25
Them	1850	0.39	Your	421	0.29	My	643	0.22
One	1636	0.26	My	402	0.28	Them	635	0.22
Your	1472	0.23	They	395	0.28	<u>He</u>	575	0.20
Which	1283	0.20	Something	297	0.21	One	574	0.19
<u>He</u>	1095	0.17	Them	290	0.20	Which	568	0.19
Something	939	0.15	Me	256	0.18	Me	477	0.16
My	920	0.14	Which	240	0.17	<i>Our</i>	463	0.16
Their	907	0.14	<i>Us</i>	186	0.13	Something	334	0.11
<i>Us</i>	876	0.14	<i>Our</i>	161	0.11	<i>Us</i>	307	0.10
Me	738	0.12	Anything	116	0.08	Everyone	215	0.07

An LL comparison of the use of *we* in the Business data, compared to the other data sets illustrates that significant difference in frequency exists, with scores of +637.12 when compared to Bioinformatics and +404.57 when compared to Marine Engineering. There was no significant difference in the use of *we* across the latter two of these data sets; however, nor was there any significant difference in the use of *I* across them all. In terms of semantic association, the proliferation of this pronoun and the other first person plural pronouns in the Business data functions to create a sense of group membership and inclusivity. This is something that is further enhanced by the frequent use of the determiner *our* in the Business data, occurring with the relative frequency of 0.54, compared with 0.16 for Bioinformatics and 0.11 for Marine Engineering, so with LL scores of +622.48, +357.68; +834.53 and +299.69 respectively.

Beyond pronouns, we see a slight difference in the ranking of the word *yeah* (italicized in Table 11.2) across each of the subjects, occurring at a

relative frequency of 1.63 in the Business data, 1.97 in Bioinformatics and 2.55 in Marine Engineering. *Yeah* is classified as an item which features in the ‘Discourse Bin’ in *WMatrix*. Amongst other things, these items include discourse markers and emphatic communication terms. Discourse markers, specifically, are high-frequency items that are typically more characteristic of spoken than written communication (Aijmer 2004) and are considered to have little or no propositional meaning (Fung and Carter 2007), but are multifunctional (Schiffrin 1987: 106), with their function being highly dependent on the context in which they are used. They ‘not only organize the discourse but can indicate degrees of formality and people’s feelings towards the interaction’ (Carter and McCarthy 2006: 212), and therefore often function as interactional features. Consequently, the presence of discourse markers, and other items in the ‘discourse bin’, offers a useful way into the CA analysis (see below), providing a focus and direction for a more fine-grained investigation of specific features of the interaction. In short, identifying discourse markers and other interactional features tells researchers ‘where to look’ in subsequent CA analyses.

A list of the top 10 most frequently used ‘discourse bin’ words and phrases (which are, again, not simply limited to discourse markers) across the subjects is shown in Table 11.4.

Although somewhat variable from one subject to the next in terms of rank, frequency and relative frequency, there is no significant statistical

Table 11.4 The most frequently used ‘discourse bin’ terms featured in the Business, Bioinformatics and Marine Engineering data

Business			Bioinformatics			Marine Engineering		
Word	Freq.	Rel.	Word	Freq.	Rel.	Word	Freq.	Rel.
Yeah	10242	1.63	Yeah	2813	1.97	Yeah	7510	2.55
Er	3675	0.58	Oh	777	0.54	Erm	1929	0.65
Um	2628	0.42	No	707	0.49	Eh	1187	0.40
Erm	2483	0.40	Er	657	0.46	No	1180	0.40
No	1795	0.29	Erm	492	0.34	Er	1172	0.35
Eh	1772	0.28	Uh	400	0.28	Um	1039	0.27
Oh	1583	0.25	Eh	366	0.26	Right	810	0.27
Mm	1120	0.18	Right	361	0.25	I mean	808	0.27
Right	1070	0.17	Ah	320	0.22	Uh	805	0.27
Uh	1050	0.17	Um	319	0.22	I think	681	0.23

difference in the frequency of use of the individual items shown in Table 11.4. Collectively, the data reveal that such discursive items actually occur more frequently in the Bioinformatics data than in the other subjects. The most significant difference in use is between this subject and the Business data, where these items occur with a relative frequency of 7.08 and 6.18 in the latter, with an LL score of +144.38. These items are also more prevalent in the Marine Engineering data than in the Business data, occurring 22,335 times (relative frequency of 7.58) in this data set, with a LL of +573.27. There was no significant difference in the use of these terms across the Bioinformatics and Marine Engineering data, however. This suggests that the speech in the Business data is perhaps more formal and more transactional than interactional. This is particularly interesting in light of the frequent use of the collective rather than personal pronouns discussed above, as it provides the sense of a collective rather than individualistic voice in the speech—more group-led than personal.

While these discursive items proliferated in the three data sets, grammatical items (that is, function rather than content words, such as prepositions, adverbs, conjunctions and so on) are statistically more frequent in the Business data than in the Bioinformatics data, with relative frequencies of use at 25.18 for Business and 23.74 for Bioinformatics (with an LL score of +97.86). However, the rate of use in the Marine Engineering data was more similar to the Business data in this respect, with a relative frequency of 25 for the use of these items, with Business scoring an LL of only +2.42 when compared to these data. There is no significant difference in the rate of use of grammatical items across the Bioinformatics and Marine Engineering data, however. Grammatical items are functional units in language, so their proliferation in the Business data again hints towards the more transactional nature of this discourse.

Thus far, our CL analysis has provided some interesting insights into our corpus concerning the use of pronouns and discourse items. However, in order to gain a deeper understanding of spoken interactions in different educational settings, we needed to see how the salient features identified using CL actually operated in speakers' turns and in longer sequences of interaction. It is at this point that another framework was needed: by looking at microcontexts within a CA framework, we were able to bring their interactional and pedagogical relevance into relief (as

we detail below). The dialectic between CA and CL thus allowed us to better understand why certain items were clustering at certain points.

3.4 CA Analysis

Data were transcribed using a notation system derived from the work of Gail Jefferson (2005) and then analysed following the principles and theoretical underpinnings of CA, an approach that ‘has evolved from ethnomethodology’ (Kasper and Wagner 2011: 117). Its aim is to offer a fine-grained and emic description of naturally occurring spoken data as a means of understanding ‘talk as a basic and constitutive feature of human social life’ (Sidnell 2010: 1). In a summary of the literature, Seedhouse (2005) identifies the following basic principles of CA: there is order at all points—all conversations are highly structured and ordered; contributions are context-shaped and context-renewing: any one contribution is both shaped by and shapes the context in which it occurs, which means that any understanding of turns-at-talk can only take place by reference to the sequential environment in which they occur; data should be approached from a bottom-up perspective without preconceptions. Essentially, the aim is to see the interaction through the eyes of the interactants. Our main rationale for choosing a CA-informed methodology is its focus on the details of the talk, enabling us to examine key aspects of educational discourse such as the relationship between specific interactional features and learning, or the extent to which the discourse creates space for learning.

In the data analysed in the present study, a number of features of the discourse were examined from a CA perspective as a means of understanding the ways in which orientations to learning and knowledge emerge across each of the three subject areas (Business, Bioinformatics and Marine Engineering). One of our main points of interest was the nature of turn-taking, a feature which lies at the very heart of CA (Hutchby and Wooffitt 2008). Other features which attracted our attention include repair, pausing, adjacency pairs, topic management, participation rights and preference structure (though not all of these features are reported here).

Fig. 11.1 shows an extract taken from a group activity which involves seven students, working independently of their tutor, in a business school

1	<\$4>	I'm just worrying what's going to happen (1.0) if we reduce
2		<their> (.) wage (0.8) at Christmas
3		(0.2)
4	<\$1>	the problem is (0.2) if (.) they strike over Christmas↓ that's
5		.hh pretty .hhihi much .hh us screwed .hhh
6		(0.7)
7	<\$5>	I think keep it as six ninety
8		(0.5)
9	<\$6>	but we're still going to be higher than other people's
10		wage↓ >they're not going to leave are they?<
11		(.)
12	<\$4>	[I don't think they'll <u>strike</u>]
13	<\$1>	[Yeah but they may might] change ° theirs°
14		(0.5)
15	<\$4>	They won't I don't think they would strike because↓ (2.3)
16		they've got nowhere else to go
17		(0.8)
18	<\$5>	yeah. (.) [.hh HAhaha↓
19	<\$4>	[.hhh hahaha =
20	<\$6>	= But they might work less efficiency (.)° efficiently°
21		(0.5)

Fig. 11.1 Extract from NC0030 Business game (0.54–1.27)

seminar. In the activity, the students are playing a business game in which they must manage a hypothetical company. Prior to the beginning of the extract, participants in the business game are involved in a discussion in which they consider ways of reducing wages paid to employees over the Christmas period. This discussion is followed by an extended 54-second pause. (Transcription conventions can be found in the Appendix.)

In line 1, S4 reinitiates the talk by first saying 'I'm just worrying' which indicates that they have some concerns about what will follow, and then pauses for 1.0 second. As this is not a transition relevance place (TRP: a point in any spoken encounter where another speaker has the possibility of taking a turn), no one takes up the floor during this pause. Following a 1.0 second pause, S4 continues their turn stating their main concerns. There is some evidence in lines 1–3 of the use of hesitation devices, signalled by the pauses and micropauses and by the slowing down of the word 'their' (<their>). This hesitant delivery continues in lines 4–6,

as evidenced by a further use of pausing and micropausing and breath intakes (.hh) by student 1. Students 4 and 1 are essentially ‘feeling their way’ in the discourse and coming to a realization that any reduction in wages is potentially problematic and should be avoided, evidenced in line 7 by student 5’s evaluation ‘I think keep it as six ninety’ (that is, keep the hourly wage at £6.90). Following a 0.5 second pause in line 8, student 6’s second-pair part offers a possible reason—higher wages—for workers not to leave.

The micropause in line 11 is followed by a stepwise topic change (Sacks 1992), in which student 4 (in line 12) responds to speaker 1’s assertion that the workers might strike. In order to understand what is ‘really happening’ in this encounter then, CA allows us to relate student 4’s turn in line 12 to student 1’s turn in line 4. The turn in line 12 is essentially a second-pair part to the initial topic launch in line 4. Clearly, our CL analysis is unable to uncover such intricacies of turn-taking, floor management and topic control. And yet, we suggest, this layer of analysis is crucial to making sense of the encounter.

The overlapping speech in lines 12 and 13 is also of interest. At first glance, it would appear that these two turns (student 4’s assertion [I don’t think they’ll strike], said in overlap with student 1’s claim [Yeah but they may might] change ° theirs °) are sequential. In other words, line 13 is the second-pair part of line 12. On closer examination, however, it is apparent that student 1’s turn in line 13 actually refers back to lines 9 and 10 where student 6 makes the observation ‘but we we’re still going to be higher than other people’s wage ↓ >they’re not going to leave are they? <’. The difficulty in the analysis stems in part from the use of the pronoun ‘they’ and the personal pronoun ‘theirs’, said softly in line 13 (° theirs °). It is CA’s focus on turn sequence and adjacency pairs which allows us to relate the utterance produced in line 13 to that of student 6 in line 9.

In addition to a pure CA analysis of this extract, we can make some important observations concerning the use of pronouns, linking back to the previous section where CL was used to study the same feature. In the opening lines of the extract, S4 initially uses the first person singular *I*, which demonstrates that what will follow is her ‘personal’ concern; lines 1–2 offer a personal assessment of a business situation. Subsequently, however, they say ‘if *we* reduce *their* wage’. Generally in the business

game, the use of *we* might refer to either the participants in the business game as students, or managers of the company in the game. In line 2, *their* refers to the workers in the company and, as such, it could be claimed that *we* in line 1 refers to the participants in the game as ‘managers of the company’.

Following line 2, there is a 0.2 second pause, which is then followed by S1 taking the floor. In lines 4 and 5, S1 responds to the ‘concern’ stated by S4 in previous lines. While responding, S1 also uses personal pronouns in a similar way, referring to the participants in the business game as the company management team and using *us* to refer to themselves in opposition to *they*, meaning the workers. This finding stands in contrast to our previous observations (see above) concerning the use of personal pronouns in an educational setting, where they both indicate and promote a sense of inclusivity, belonging and space for learning. Here, we see the opposite, a sense of ‘othering’ (Said 1995) in which management and workers are in some kind of opposition and where there is a need to delineate management actions (reducing hourly pay) from possible resultant worker reactions (a strike).

In line 9, S6 takes the floor following a 0.5 second pause. In this turn, S6 makes a comparison between their workers’ wages and other people’s wages by suggesting, ‘we’re still going to be higher than other people’s wage↓’. In this turn, S6’s use of *we* might refer to participants’ company management team and/or students at the same time as there is a comparison between ‘*other people’s wage*’. ‘Other people’ in this turn might also index other groups in the business game as students working together or other ‘company management teams’. In line 10, the workers are referred to as ‘*they*’. After a micropause in line 11, S4 agrees with S6 by stating their own position using ‘I don’t think’ and they also use ‘*they*’ to refer to the workers.

To conclude our analysis of this extract, there are some interesting observations to be made concerning the next exchange, lines 15–21. Following a 0.5 second pause in line 14, student 4 takes the floor and, with a self-initiated self-repair (They won’t I don’t think they would strike), makes the assertion that the workers are unlikely to strike. The reason for this assertion is given after a lengthy 2.3 second pause and heavily falling intonation on the word ‘because’ (indicated because↓). The pause and falling

intonation probably indicate that S4 is uncertain how they will complete their assertion that strike action is unlikely. When they eventually give their reason in line 16 (they've got nowhere else to go) there is a 0.8 second pause in line 17, followed by laughter, produced in overlap by S5 and S4 (lines 18 and 19). This apparent joyous response from S5 and S4 is, however, quickly countered by S6 in line 20. Here, S6's latched turn (marked =) and including a self-initiated self-repair, offers an assessment that even if the workers do not strike, they may still work less efficiently. In other words, the problem has still not been resolved.

This extract demonstrates that during a business game, participants use *we* and *they* to co-construct two different social groups, or 'communities of practice'. By referring to themselves as *we* and *us* constantly, the participants co-construct themselves as a 'team' with a shared aim and interest in opposition to the 'other' community of practice, the 'workers'. The management team constantly reinforce this sense of 'us and them' by referring to the workers as 'they', thereby co-constructing the workers as a homogeneous group. Through the use of such personal pronouns, participants also perform membership categorization to two communities of practice: the company management team, and workers in the company. What this analysis highlights very clearly is the extent to which a complementary CA analysis of a data set provides much greater insights than those obtained by a CL analysis alone (see discussion below).

We turn now to our second illustrative extract (Fig. 11.2). Here, the same group of students is discussing their business plan with their tutor.

1	<\$7>	At the same time as that are you still planning to purchase
2		market intelligence to see what other companies are doing?
3	<\$1>	Yes we think it's.
4	<\$7>	That's not gonna be something that you're gonna cut back
5		on.
6	<\$4>	No cos I think it's very important.
7	<\$3>	Yeah we think it's extremely important to know what
8		our competitors are doing otherwise how else would
9		we know where we're standing?
10	<\$7>	Yeah. Absolutely.
11	<\$3>	Yeah.

Fig. 11.2 Extract from NC0030 Business game

In lines 1 and 2, S5 asks a question regarding the business plan made by the group members: 'are you still planning to purchase market intelligence to see what other companies are doing?' While formulating the question, S7, the tutor, uses 'you' which might index the group members as the 'company management team', a group of students responding to a tutor's question, or only one member of the group to whom the question is directed. The second-pair part of this question-answer adjacency pair in line 3 indicates that S1 demonstrates an understanding of 'you' in line 1 as indexing the group members as she uses 'we' to answer this question. The use of 'we think' is very prevalent in the Business School data, especially when the students are presenting their business plans to other students and lecturers. It is a feature which is unique to this data set and which we might usefully investigate using a CL analysis. By identifying this feature in longer exchanges and noting its uniqueness to this context, we have the potential to see how it behaves in collocation with verbs such as *think*, *argue*, *believe* and so on, in and across NUCASE data using a CL analysis.

Returning now to our CA analysis of this extract, S7 pursues his line of questioning in lines 4–5, extending the subtopic of 'purchasing market intelligence' in the statement 'That's not gonna be something that you're gonna cut back on'. This effectively functions to give students the answer to the question from lines 1–2, by the way his/her turn is formulated, as it is unlikely that students would counterargue and say that they are going to cut back on this. Lines 1–5 follow the classic IRF (Initiation, Response, Follow-up) exchange structure which is typical of most classroom interaction. A tutor's question is met with a student's response, which is then followed up by the tutor in a further evaluation of the response. In this instance, the follow-up is both an acknowledgement of the student response in line 3 and an invitation for students to comment further on their decision, which they do in lines 6–9. S4's assessment ('No cos I think it's very important') in line 6 is taken up and extended by S3 in lines 7–9. S3's preference to agreement ('Yeah we think it's extremely important') is followed by a reason in the form of a rhetorical question. This turn is met with emphatic agreement by both the tutor in line 10 and S3 in line 11.

In terms of pronoun use, in line 6, S4's response 'No cos I think it's very important' results in a change in pronoun from *we* to *I*, signalling that they are offering a personal opinion here and that this may not actually be the position of the entire management group. In the following line, S3 responds to S4's assessment with an agreement token *yeah*, but uses 'we think' after the agreement token. The change of personal pronouns in this adjacency pair in lines 6 and 7 is very interesting. While agreeing with S4's assessment, the turn in line 7 might be interpreted as 'repairing' the use of 'I think' in line 6. By doing so, S3 is claiming a group ownership for the assessment made by S4 in line 6, and following that they are also upgrading the assessment by suggesting that it (market intelligence) is '*extremely* important', which is then followed by a question addressed to S5: 'otherwise how else would we know where we're standing?' This question is again formulated to reflect the group membership through the use of 'we'. Also, by redirecting the question to S5, S3 makes it relevant for S5 to state their own stance regarding the answer given by the group members. S5's answer in line 10 initially begins with an agreement token 'yeah' and then upgraded by 'absolutely' to indicate a strong alignment with the position taken by the group members.

In addition to offering a finer-grained investigation of our Business data, the CA analysis sheds extra light on the findings of the CL study in the previous section. The latter demonstrated that the frequency of *we* usage was significantly higher in the Business data. The analyses above show in fine detail how *we* and *they* might be used in Business classes. In the Business School data, the use of business games is a very common activity, allowing students to simulate real-life business decision-making and deal with problems as they arise. Our two extracts demonstrate, in some detail, how students engage with the activity and successfully manage two different participation frameworks at the same time: (a) students taking part in a business game; (b) their role-play identity as members of a company management team. As such, they use *we* very frequently during the games to establish themselves as members of the same group and to organize decision-making. In these games, students need to make various decisions about the company and in the process of decision-making, they use *we* all the time.

4 Discussion

This chapter set out to combine CL with CA to provide enhanced descriptions of spoken interaction in a SGT HE context. Our aim was to identify which discursive features were characteristic of three specific educational contexts (Business, Bioinformatics and Marine Engineering) and to consider how these features varied in use from one context to another. Our methodology entailed the use of two distinct analytical approaches (CL and CA) with the NUCASE educational discourse corpus of 1 million words.

Our results allowed us to make comparisons both within and across these three interactional contexts. For example, when we compare business talk with both bioinformatics and marine engineering very different profiles or ‘fingerprints’ (Heritage and Greatbatch 1991) emerge. In the Business data, pronouns appeared to do much more interactional work than in the other two subcorpora, often indicating association or otherwise with other group members, highlighting the importance of inclusivity in educational settings and demonstrating ways in which interactional space might be opened up or closed down. In the same data set, particular pronouns and discourse markers were used to delineate space for learning and to mark membership of specific groups. One of the main attractions of this combined CLCA methodology, we suggest, is that it frequently operates at the interface of individual words, spoken utterances and text. The interdependence of CL and CA in the methodology is, to a large extent, mirrored in its main enterprise: here, the ways in which words and text combine to co-construct meanings. By looking first at the ‘big picture’ of the text and by using a CL analysis, we identified particular foci which could then be investigated in more detail. In this study, data concerning the range, type, frequency and collocation of discourse markers helped us to identify specific features in the text to study in greater depth using CA. This approach has particular relevance to the study of spoken discourse where so much meaning is conveyed by features which are not apparent in a CL analysis. Our combined CLCA analysis enabled us to make much more reliable assertions about what is really ‘happening’ in spoken discourse and revealed some of the mechanisms by which

interactants create and demonstrate group membership, orient towards a particular perspective or opinion, reveal different participation structures and so on. More importantly, in an education setting, this approach enables researchers to make more reliable claims about the highly complex interplay of word, utterance and text in relation to orientations to learning and knowledge.

Increasingly, small, highly contextualized corpora are being used to advance understandings of spoken discourse. As a mixed method, we believe that CLCA combines quantitative and qualitative approaches to data analysis very effectively. Perhaps its most important attribute is the iterative manner in which the methodology approaches data. There is a genuine and principled movement from a quantitative to qualitative analysis and back again, ensuring that the two methodologies are brought into close relief. In much mixed methods research in applied linguistics and social sciences, this is not the case; how often do we read reports using, for example, questionnaires and interviews, where the results from each method are presented separately and where there is little or no attempt to show how the findings from one support or refute the other? In a CLCA methodology, each data set informs the other and enables a more principled approach to analysis.

The NUCASE corpus has huge potential in the context of HE professional development. For example, it could be used to help tutors promote more engaged, dialogic teaching environments and develop fine-grained understandings of their approach to teaching, especially if some of the methodological tools employed in the study are used or adapted. A key element of continuing professional development (CPD) is the capacity to reflect on practice, a process which is greatly facilitated when lecturers have access to appropriate tools and a way of collecting and analysing data. The NUCASE corpus, we suggest, might be used to promote reflection and engage tutors in discussions about their teaching performance, ideally with a peer or 'critical friend'. There have been recent calls (see, for example, Mann and Walsh 2013) for reflective practice to be repositioned and reinvigorated by making the process data-led, dialogic and through the use of appropriate tools. A corpus such as NUCASE has the potential to facilitate this change: reflections by university tutors should, and indeed could, be enhanced through the

use of actual recordings and the kinds of tools for analysis described in this chapter.

The corpus could also be used by educators working in, for example, staff development units in universities, whose main concern is the enhancement of teaching. In the current market-led and client-oriented context in which those of us working in UK HE must now operate, teaching quality is of concern to everyone. In order to enhance quality and help lecturers develop their practice in a meaningful and evidence-based way, the use of corpora such as NUCASE has much to offer. It could, for example, form one element of a stimulated recall session in which a more experienced practitioner engages in a professional dialogue with more junior colleagues using video playback and/or a transcript. By looking more closely at actual recordings and transcriptions of them, practitioners could develop more detailed understandings of, for example, their use of open or closed questions; their ability to engage learners in discussion and debate, and their skill in promoting critical thinking, one of the hallmarks of HE teaching and learning.

We have already mentioned the capacity of NUCASE to enhance our understandings of appropriate admissions criteria, especially in relation to the English language requirement. There is scope too to develop closer understandings of the materials used in seminars and tutorials. Which tasks generate the most discussion, for example? How do students respond to different types of activities? How do we judge levels of participation and engagement? These features are crucial to enhancing understanding of teaching and learning in HE—a corpus-based approach to both professional development and materials design enables much finer-grained understandings to emerge.

Moving beyond educational settings, we suggest that a CLCA methodology has value in a broad range of professional contexts where ‘up-close’ and detailed understandings of interactions are required. For example, in doctor–patient interactions, doctors in training need to understand the finer nuances of their interactions with patients if they are to ‘get the story’ and recommend treatment (see Hesson et al. 2012). By using a CLCA methodology, it would be possible to demonstrate how certain linguistic and interactional features are better suited to certain microcontexts than others. For example, when eliciting from a patient, doctors need to pay

attention to both social and biomedical information. A CLCA analysis of transcripts would demonstrate quite clearly how this information is used in making a diagnosis. More importantly, such an analysis would enable doctors to reflect on their professional practices and make changes in light of the linguistic evidence presented.

Consider too contexts where experts from different professions come together, as is often the case, to solve a problem or set up a new project. It is very common, for example, to find engineers, architects, designers and project managers working together. In such cross-disciplinary settings, multiple discourses are used, often with different meanings, with the result that misunderstandings and communication breakdowns are frequent. A CLCA analysis of such interactions may uncover some of the reasons for misunderstandings. A key technical term, for example, may have diverse meanings in different communities of practice. Pronoun use, as illustrated in this chapter, is important to developing understandings of inclusion or exclusion—what CDA terms ‘centring’ or ‘othering’. Essentially, pronoun use can indicate whether a person is showing allegiance or opposition to an idea or concept. Consider too the ways in which a CLCA analysis may help to foster understandings of interactional competence in different professional settings. At the very minimum, it would be possible to develop closer understandings of the ways in which people in different workplace settings co-construct meanings in order to make sense of their professional worlds in the pursuit of a common institutional goal (see also Vine, Chap. 12 for further exposition of this idea).

5 Conclusion

In this chapter, we have offered a brief exposition of how CL might usefully combine with CA in the description and analysis of spoken interactions in an HE SGT context. Using data from a 1 million word corpus, NUCASE, we have tried to demonstrate the value of using a combined CLCA methodology. Had we used CL on its own we would have achieved interesting lists of high-frequency items which we could have explained functionally but it would not have brought us anywhere near the depth

of understanding compared with what a CA framework could explain. Had we looked at the data purely from a CA perspective, we would have possibly identified a number of interesting phenomena across the corpus, but we would not have been able to support the fact that the words and patterns they contain were actually high-frequency items (that is, keywords and high-frequency words). By combining two seemingly mutually exclusive methodologies, we were able to demonstrate the ways in which words, utterances and texts combine to construct meaning. It would seem safe to say then that CL and CA have much to offer one another and that there is huge potential for this kind of research, especially where small, locally derived, context-specific corpora are being used.

Looking towards the future, it seems highly likely that some of the technological advances made in CL will influence developments in CA. Studies are already underway to tag turn openings and closings in order to provide larger-scale evidence concerning how turns are opened, closed and switched. It is not inconceivable that other features of CA might be treated in a more quantitative way [pauses perhaps, as demonstrated phonetically in Heldner and Edlund (2010) for Dutch, Swedish and Scottish English]. There is certainly scope for CA to make more and more detailed references to the interactional work performed by individual words, a point already made by one of the best-known proponents of CA, Heritage (2012). In the same vein, it is interesting to speculate how CL might develop a closer relationship with CA, using some of its mechanisms and procedures. It would not be difficult to imagine how CA principles could be extended to the CL operation of concordancing, especially in cases where concordance lines operate across turns. Similarly, by looking more closely at content words, CL could offer useful insights into the ways in which topics are established, developed and maintained in spoken discourse; topic is, of course, another of CA's concerns and yet one which, in recent years, has been under-researched.

There is then much to be gained and little to lose by experimenting with a combined CLCA methodology. We would encourage researchers working in a CL paradigm to experiment with a combined CLCA approach, not necessarily in the way described here, but in a structured, principled and iterative way which is likely to be of benefit to other researchers as well as non-academic audiences.

6 Appendix: Transcription Conventions [Adapted from Hutchby and Wooffitt (2008)]

- (1.8) Numbers enclosed in parentheses indicate a pause. The number represents the number of seconds of duration of the pause, to one decimal place. A pause of less than 0.2 seconds is marked by (.)
- [] Brackets around portions of utterances show that those portions overlap with a portion of another speaker's utterance.
- = An equals sign is used to show that there is no time lapse between the portions connected by the equal signs. This is used where a second speaker begins their utterance just at the moment when the first speaker finishes; a *latched* turn.
- :: A colon after a vowel or a word is used to show that the sound is extended. The number of colons shows the length of the extension.
- (hm, hh) These are onomatopoeic representations of the audible exhalation of air
- .hh This indicates an audible inhalation of air, for example, as a gasp. The more h's, the longer the in-breath.
- ? A question mark indicates that there is slightly rising intonation.
- . A period indicates that there is slightly falling intonation.
- , A comma indicates a continuation of tone.
- A dash indicates an abrupt cut off, where the speaker stopped speaking suddenly.
- ↑↓ Up or down arrows are used to indicate that there is sharply rising or falling intonation. The arrow is placed just before the syllable in which the change in intonation occurs.
- Under Underlines indicate speaker emphasis on the underlined portion of the word.
- CAPS Capital letters indicate that the speaker spoke the capitalized portion of the utterance at a higher volume than the speaker's normal volume.

- This indicates an utterance that is much softer than the normal speech of the speaker. This symbol will appear at the beginning and at the end of the utterance in question.
- > <, < > ‘Greater than’ and ‘less than’ signs indicate that the talk they surround was noticeably faster, or slower than the surrounding talk.
- (would) When a word appears in parentheses, it indicates that the transcriber has guessed as to what was said, because it was indecipherable on the tape. If the transcriber was unable to guess as to what was said, nothing appears within the parentheses.
- £C'mon£ Sterling signs are used to indicate a smiley or jokey voice.

References

- Ädel, Annelie. 2010. How to use corpus linguistics in the study of political discourse. In *The Routledge Handbook of Corpus Linguistics*, eds. Anne O’Keeffe, and Michael McCarthy, 591–604. London: Routledge.
- Aijmer, Karin. 2004. Interjections in a contrastive perspective. In *Emotion in Dialogic Interaction*, ed. Edda Weigand, 99–120. Amsterdam: John Benjamins.
- Atkins, Sarah R. 2011. A Cognitive Linguistic Perspective on Social Space in Online Health Communities. PhD dissertation, University of Nottingham.
- Basturkmen, Helen. 2002. Negotiating meaning in seminar-type discussions and EAP. *English for Specific Purposes* 21(1): 233–242.
- Bennett, Catherine, Christine Howe, and Emma Truswell. 2002. *Small Group Teaching and Learning in Psychology*. York: LTSN Psychology University of York.
- Benwell, Bethan M., and Elizabeth H. Stokoe. 2002. Constructing discussion tasks in university tutorials: Shifting dynamics and identities. *Discourse Studies* 4(4): 429–453.
- Biber, Douglas. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes* 15: 133–163.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.

- Carter, Ronald A., and Michael J. McCarthy. 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press.
- Chafe, Wallace L., and Jane Danielewicz. 1987. Properties of spoken and written language. In *Comprehending Oral and Written Language*, eds. Rosalind Horowitz, and S. Jay Samuels, 83–113. New York: Academic Press.
- Cotterill, Janet. 2010. How to use corpus linguistics in Forensic Linguistics. In *The Routledge Handbook of Corpus Linguistics*, eds. Anne O’Keeffe, and Michael M. McCarthy, 578–590. London: Routledge.
- Fung, Loretta, and Ronald Carter. 2007. Discourse markers and spoken English: Native and learner use in pedagogical settings. *Applied Linguistics* 28(3): 410–439.
- Gibson, Will, Andy Hall, and Peter Callery. 2006. Topicality and the structure of interactive talk in face-to-face seminar discussions: Implications for research in distributed learning media. *British Educational Research Journal* 32(1): 77–94.
- Heldner, Mattias, and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38: 555–568.
- Heritage, John. 2012. Epistemics in action: Action formation and territories of knowledge. *Research on Language and Social Interaction* 45(1): 1–29.
- Heritage, John, and David Greatbatch. 1991. On the institutional character of Institutional Talk: The case of news interviews. In *Talk and Social Structure: Studies in Ethnomethodology and Conversation Analysis*, eds. Deirdre Boden, and Don H. Zimmerman, 93–137. Berkeley: University of California Press.
- Hesson, Ashley M., Issidoros Sarinopoulos, Richard M. Frankel, and Robert C. Smith. 2012. A linguistic comparison of patient- and doctor-centered interviewing. *Patient Education and Counseling* 88: 373–380.
- Heylighen, Francis, and Jean-Marc Dewaele. 2003. Variation in the contextual-ity of language: An empirical measure. *Foundations of Science* 7: 293–340.
- Hutchby, Ian, and Robin Wooffitt. 2008. *Conversation Analysis*, 2 edn. Cambridge: Polity Press.
- Jefferson, Gail. 2005. Glossary of transcript symbols with an introduction. In *Conversation Analysis: Studies From the First Generation*, ed. Gene H. Lerner, 13–31. Amsterdam: John Benjamins.
- Kasper, Gabrielle, and Johannes Wagner. 2011. A conversation-analytic approach to second language acquisition. In *Alternative Approaches to Second Language Acquisition*, ed. Dwight Atkinson, 117–142. London: Routledge.
- Knight, Dawn, Svenja Adolphs, and Ronald Carter. 2013. Formality in digital discourse: A study of hedging in CANELC. In *Yearbook of Corpus Linguistics and Pragmatics*, ed. Jesús Romero-Trillo, 131–152. London: Springer.

- . 2014. CANELC: Constructing an e-language corpus. *Corpora* 7(1): 29–56.
- Koester, Almut. 2006. *Investigating Workplace Discourse*. London: Routledge.
- Levinson, Stephen. 2006. Cognition at the heart of human interaction. *Discourse Studies* 8(1): 85–93.
- Lorés, Rosa. 2006. The referential function of metadiscourse: Thing(s) and idea(s) in academic lectures. In *Corpus Linguistics: Applications for the Study of English*, eds. Ana María Hornero, María José Luzón, and Silvia Murillo, 315–334. Bern: Peter Lang.
- Louwerse, Max M., Scott A. Crossley, and Patrick Jeuniaux. 2008. What if? Conditionals in educational registers. *Linguistics and Education* 19(1): 56–69.
- Mann, Steve, and Steve Walsh. 2013. RP or ‘RIP’: A critical perspective on reflective practice. *Applied Linguistics Review* 4(2): 291–315.
- O’Keeffe, Anne. 2006. *Investigating Media Discourse*. London: Routledge.
- O’Keeffe, Anne, and Fiona Farr. 2003. Using language corpora in language teacher education: Pedagogic, linguistic and cultural insights. *TESOL Quarterly* 37(3): 389–418.
- O’Keeffe, Anne, and Michael McCarthy, eds. 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.
- Rayson, Paul E. 2003. Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison. PhD dissertation, Lancaster University.
- Sacks, Harvey. 1992. Stepwise topical movement. In *Lectures on Conversation*, vol 2, ed. Harvey Sacks, 291–302. Cambridge, MA: Blackwell.
- Said, Edward. 1995. *Orientalism*. London: Penguin Books (first published in 1978).
- Santamaría-Gracia, Carmen. 2010. Bricolage assembling: CL, CA and DA to explore agreement. *International Journal of Corpus Linguistics* 16(3): 345–370.
- Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.
- Seedhouse, Paul. 2004. *The Interactional Architecture of the Second Language Classroom*. London: Blackwell.
- . 2005. Conversation Analysis and language learning. *Language Teaching* 38(4): 165–187.
- Sidnell, Jack. 2010. *Conversation Analysis: An Introduction*. West Sussex: Wiley-Blackwell.
- Simpson, Rita, Sarah Briggs, Janine Ovens, and John Swales. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

- Stokoe, Elizabeth H. 2000. Constructing topicality in university students' small-group discussion: A conversation analytic approach. *Language and Education* 14(3): 184–203.
- Thompson, Paul, and Hilary Nesi. 2001. The *British Academic Spoken English* (BASE) Corpus Project. *Language Teaching Research* 5(3): 263–264.
- Viechnicki, Gail B. 1997. An empirical analysis of participant intentions: discourse in a graduate seminar. *Language and Communication* 17(2): 103–131.
- Walsh, Steve. 2011. *Exploring Classroom Discourse*. London: Routledge.
- Walsh, Steve, and Anne O'Keeffe. 2007. Applying CA to a modes analysis of third-level spoken academic discourse. In *Conversation Analysis and Languages for Specific Purposes*, eds. Hugo Bowles, and Paul Seedhouse, 101–139. Bern: Peter Lang.
- Wilson, Andrew, and Paul Rayson. 1993. Automatic content analysis of spoken discourse: A report on work in progress. In *Corpus-based Computational Linguistics*, eds. Clive Souter, and Eric Atwell, 215–227. Amsterdam: Rodopi.

12

The Wellington Language in the Workplace Project: Engaging with the Research and Wider Communities

Bernadette Vine

1 Introduction

The workplace is a domain where communication issues are of crucial importance to everyone: whether you are a worker or a member of the public, work talk has relevance for us all. Effective workplace communication enables workplaces to be productive and provide a rewarding environment in which to work and, depending on the type of workplace, for members of the public to access what they require in a stress-free way. Research on workplace communication has the potential to provide valuable input for workplaces to achieve these goals, and dissemination of research results beyond normal academic outlets must be an important consideration.

This chapter provides an overview of the Wellington *Language in the Workplace Project* (LWP) and explores a number of ways that this project shares its research in non-academic forums and uses it to benefit a wide range of people. Although not being perceived in these terms, the project

B. Vine (✉)

Victoria University of Wellington, Wellington, New Zealand

has from the outset adopted a community-based approach (Czaykowska-Higgins 2009), and this is reflected in the design, methodology and approaches to knowledge dissemination. The LWP was specifically conceived as both academic and applied with an emphasis on collaboration, building relationships and attention to the needs and concerns of the communities being researched (Stubbe 2001). In Sect. 2 below, the LWP database and basic methodology are outlined, describing the project's objectives and highlighting the diverse range of workplace contexts investigated, as well as the approach taken.

In Sect. 3, the ways the project engages with the research community are examined more closely. Engagement with the research community begins before any data is collected through consultation with potential data sites. This process often shapes and refines our research goals. Participants are also recruited to be actively involved in data collection. An integral part of our methodology is also to provide feedback to workplaces on our analysis of their data and this includes conducting workshops and producing written reports. Engagement with the workplaces in this way has helped direct our ongoing research objectives, as well as aiding the participants in developing a fuller understanding of how they communicate and the importance of effective communication in the workplace. This, in turn, has resulted in the development of a model for communication evaluation and enhancement designed to be used by workplaces. The supplementation of workplace data with interviews and other types of data has further enriched our database and our understanding of how our research participants communicate at work.

Avenues that the project has explored in order to engage with the wider community are outlined in Sect. 4. The workplaces we have researched are part of the New Zealand business community, so their needs and concerns reflect those of the wider community. Our research has been shared with the wider business community through newsletters, non-academic journals and other media outlets, as well as through presentations, workshops and seminars, both in New Zealand and overseas. We also make our methodology and information on our data-processing procedures available for other researchers working in this area. A further extension of the research has involved producing materials based on authentic data

for teaching and for human resource programmes. We have also provided advice and resources to government agencies which support migrants and employers. These aspects of the project reach out to people not only currently working in New Zealand, but also to those who are involved in training for work or considering coming to this country.

2 Database and Basic Methodology

Between 1996 and 2014, the LWP team has compiled a corpus of approximately 2000 interactions, involving 400 hours of data and over 700 people from 35 workplaces. A diverse range of workplaces is represented, from government departments to private white-collar organizations, factories, a hospital ward, eldercare residences (that is, residential care homes) and building sites. A range of different types of interaction have been recorded in these different workplaces. In white-collar workplaces, participants audio-recorded various types of meeting, including small informal meetings and large formal ones [see Holmes and Stubbe (2003) for detailed information on methodology]. In workplaces such as factories, hospitals, eldercare residences and building sites, participants were typically wired up with radio microphones or small portable recorders and recorded a range of interactions during the course of their shifts (Stubbe 2001; Major and Holmes 2002). In the factory, this included brief interactions between workers on the packing line, as well as talk between various staff members and the team leader as she visited different parts of the factory. Builders on the building sites recorded interactions with other builders as they worked, as well as recording tea breaks and interactions with the house owner and other tradesmen who visited the site.

Different types of role relationships between participants are also represented in the different workplaces. For example, in the government departments, managers, policy analysts and administration staff were all recorded interacting with people at different levels, while in the eldercare settings, recording targeted interactions between caregivers and residents. In the hospital, where nurses were the focus participants, interactions with patients constituted the main database, although our corpus also includes interactions with doctors and other nurses.

The main objectives of our research have not changed as the project has focused on a broader range of workplaces. The initial data collection involved gathering interactions in four government workplaces with the overarching aim of identifying characteristics of successful interpersonal communication. Additional aims were to diagnose causes of miscommunication, provide feedback about our findings to the participating workplaces, and to examine the implications of the findings for the way employment relations develop, in order to provide relevant input to human resource development programmes (LWP Government Presentations 1997–1998). From the outset, therefore, there was a focus on developing collaborative relationships with workplaces and exploring ways that our research could benefit the communities being researched.

The aim of identifying characteristics of effective interpersonal communication in the workplace is a theme that has continued to underlie our research as the data collection has extended to a more diverse range of workplaces. In some cases, the research has had additional goals. Hence, in an early extension of the project beyond white-collar organizations, we examined data from a plant nursery, a recycling company and a child-care facility, workplaces which had all agreed to take on workers with intellectual disabilities. Identifying the patterns of interaction present in authentic workplace data from the workplaces where the workers would be placed enabled us to aid the newcomers in developing skills and strategies which would help them adapt and fit in (Holmes and Fillary 2000; Holmes et al. 2000). In our leadership study, which involved four workplaces, there was the added aim of examining ‘how leaders communicate in “ethnicised” organisations’ (Holmes et al. 2011: 3). In more recent work with building sites and eldercare facilities, there has been the additional goal of gathering data which can be modified to produce teaching materials of use to practitioners who are teaching and supporting refugees and migrants hoping to work within these industries (see McCallum 2013; Riddiford 2013a, b, c).

Throughout our data collection at different types of workplace, whoever the people involved and whatever the particular aims have been, we have always engaged with our research participants. Following the action research principle of ‘research on, for and with’ participants (see Cameron et al. 1992: 22), and adopting an appreciative enquiry approach in which

the focus is on what is done well (Hammond 1996), the LWP set out to develop an innovative and unique participatory methodology [see Stubbe (2001) for more detail]. The resulting approach takes into account a number of factors that Czaykowska-Higgins (2009) identifies as being components of community-based language research. This model is distinguished from other models because of the way it ‘explicitly’ acknowledges and welcomes the extent to which linguists are trained by and learn from community members in issues related to language, linguistics and culture, as well as about how to conduct research and themselves appropriately within the ‘community’ (Czaykowska-Higgins 2009: 25).

A key component of such a community-based approach is the attention paid to building collaborative relationships between researchers and the community. This has ethical as well as practical implications for how research is conducted. From an ethical viewpoint, for example, the LWP has always respected the rights of participants to be informed of our research goals and the way in which data is used. Participants can withdraw consent at any time and pseudonyms are used to protect people’s identities when results are reported (for more on the LWP and ethical considerations, see Stubbe 2001). In Sect. 3, some of the practical aspects of our community-based approach are outlined.

3 Engagement with the Workplaces Being Researched

3.1 Collaborative Approach

An important characteristic of the LWP research has always been to ‘avoid researching *on*, and instead to research *with* our participants’ (Holmes et al. 2011: 32). The recruitment of volunteers within workplaces to record a range of their everyday interactions has been a fundamental aspect of this approach (Stubbe 1998, 2001; Holmes and Stubbe 2003). Since members of our team are not generally present during recording, the workers carrying the recorders and the people they interact with have control over what, when, and how much data is recorded. Volunteers are asked to record about 4 hours of data each, but sometimes have recorded

as little as 4 minutes, and at other times over 12 hours. In workplaces where larger meetings have been recorded, these have been videoed as well as audio-recorded.

Researching with our participants has also involved consultation with workplaces at the very start of the process about the research questions that will be addressed. Some organizations had questions they were eager to have answered. For instance, the management at one factory asked us to study the communication patterns in a particularly efficient and productive team since they were keen to know the ‘secret of their success’. Other organizations have been very interested in being part of ongoing research, such as our leadership and discourse project (Holmes et al. 2011). We have also provided advice and feedback on aspects of research such as recording and transcription, as well as support for research initiatives taken by workplace partners as requested, sometimes many years after our collaboration began.

3.2 Workplace Feedback Through Workshops and Reports

A further feature of this approach to researching with our participants has involved the provision of feedback to the workplaces. Providing feedback to the workplace is an integral component of our methodology and has taken different forms depending on the type of workplace involved and the requirements of each organization.

Feedback to the government departments, for example, involved workshops focusing on a number of topics, including the way that meetings were managed. For instance, illustrating meetings with both cyclical and linear progression through topics on the agenda showed how both approaches can effectively deal with topics depending on different goals. When faced with the two options—linear or cyclical—workshop participants felt that the linear approach would be more efficient and therefore more effective. We were able to demonstrate how a linear approach may be appropriate in certain cases, such as when giving an update, but this does not always allow for a thorough exploration of an issue. A cyclical approach allows meeting participants to examine an issue from different angles and identify

aspects which may need more detailed discussion. Some points may recur, each time taking the discussion a little further. For example, a team in one meeting wanted to explore a range of possible directions for their activities over the next few months. Several possible directions were initially outlined by the manager. Discussion then ranged freely, picking up different possibilities. Some points then recurred and were discussed in more depth each time. The cyclical approach used in this case was very effective and enabled a final set of agreed, preferred objectives to be identified.

Preliminary analysis of the data collected also explored topics such as the functions of humour (Holmes 1998c) and small talk (Holmes 2000a) in the workplace, and the ways that directives are given (Stubbe and Vine 1998). Presenting this material in the form of workshops and oral and written reports elicited valuable insights from the participants involved, as well as a means of validating our interpretations of the data. Participants in the workshops also identified a number of further questions and issues which they were interested in. The issues raised included the effect of different workplace cultures on styles of communication, the relationship between language and power, and ways of dealing with gender and cultural differences in styles of interaction. The workplace feedback reinforced our developing aims, and many of these themes have continued to underlie the research that members of the project team have undertaken over the intervening years.

A further positive result of presenting workshops was the feedback participants gave us about how our findings were useful for them. They commented that the data presented made them conscious of a number of communication issues and raised their awareness about their own interaction patterns. Overall there was a great deal of positive feedback about the personal and professional benefits of being involved with the LWP.

Making our initial research accessible to the workplaces through workshops was a first step in presenting our results and also gaining insights from the participants on our interpretation of their own data. Topics of feedback in written reports have included the use of humour and small talk in the workplace, meeting management and the leadership styles of managers. In each case, examples recorded in the focus workplace are used for illustration, and the format and language are designed to be accessible to non-linguists.

For instance, the extract below is taken from a report to a private organization and uses an example from one of their meetings to illustrate the point.

Humour is another of the ways groups manage their relationships. Our study shows that humour functions in complex ways in the workplace. Its power lies in its flexibility—it can function as a disguised weapon for those who want to complain, a cushion for criticism, and perhaps more importantly a source of fun and light relief among colleagues. [...] A successful attempt at humour indicates that the speaker shares a common view with others about what is amusing, thus creating rapport. Workplace humour is often jointly constructed by members of the group, thereby allowing all participants to contribute to the humour, and show their place within the in-group.

Example 1 (Capitals indicate emphasis. Pseudonyms are used throughout)

1. Jacob: oh is this THE FINAL
2. Barry: this is THE FINAL final steering committee
3. oh Pete most probably enjoyed doing that
4. Dudley: he even sent me an e mail to reinforce it with you
5. [laughter]
6. Barry: this is THE final
7. Dudley: THE final [laughs] I'm switching off the lights and I'm leaving now
8. Barry: I'm switching off the lights and I'm leaving

Each of the three, Jacob, Barry and Dudley, adds something to the humour, showing their place within the group and the shared background knowledge they have about the long awaited 'final' steering committee meeting. Also, by echoing the comments of others (as in lines 7 and 8), Barry and Dudley are showing that they understand and approve of what is funny.

Our engagement with the workplaces where we conducted the research, and the workshops and reports we produced for them, led another government department, which we had not worked with, to approach us for practical help. A person who had worked in one of the workplaces where we had initially gathered data had moved to this organization. The team she was part of knew that their meetings were not working effectively, but were finding it difficult to identify why, and to put any strategies in place

that would help them operate more smoothly. Analysis of the meetings, drawing on the research we had already carried out in New Zealand government departments, as well as non-governmental organizations (NGOs), highlighted a number of aspects where meeting processes were not working effectively to meet their goals. For example, topics on the agenda were almost always worked through in a linear fashion. This gave the impression that the participants in the meeting were not being consulted and their opinions were not being elicited. The meeting chair quickly dealt with each topic on the agenda leaving no time for exploration or discussion of issues. We provided this workplace with a report that highlighted a number of aspects of their meetings which suggested why the meetings were not running smoothly and we identified some communication strategies which should help their meetings to run more efficiently.

3.3 Communication Evaluation and Development Model

The positive feedback that we had from participants in our research about the benefits to them led to the development of a Communication Evaluation and Development (CED) model by members of the LWP team (see Jones and Stubbe 2004). From discussions with workplaces, it was clear that traditional methods for communication assessment and training available to workplaces failed to deliver. These methods tend to ‘focus on a set of discrete, rigidly defined communication skills or tasks, which it is assumed can be learned in isolation and readily transferred into different workplace contexts’ (Jones and Stubbe 2004: 189). From the beginning, our feedback to workplaces had highlighted the fact that effective communication involved flexible strategies which varied depending on a range of contextual factors. It was also apparent that using real interactions from a workplace had much more impact when conducting communication training, as did reaching an understanding with the parties involved of the issues that were important.

These insights were developed into a CED model which could be used by workplaces. A trial of the model was carried out in one organization

where we had previously collected data. This involved two workers who were encouraged to reflect on their communication and plan how they might modify the strategies they used to communicate. Both participants identified a communication issue that was problematic for them, and through the process of applying the model and reflecting on their communication felt that they were able to make positive changes. The LWP team members involved in this project observed that the development of ‘the ability to observe, analyse and reflect on the communicative challenges in one’s own particular work environment, and then to initiate cycles of change, is empowering for the individual’ (Jones and Stubbe 2004: 204).

3.4 Supplementing Workplace Data

As mentioned above, our core methodology has involved recruiting volunteers from each workplace to record a range of interactions. The leadership project we undertook, however, added another aspect with the addition of interviews. This project was undertaken in collaboration with Brad Jackson, then Director of the Centre for the Study of Leadership and Head of the School of Management at Victoria University of Wellington. He was keen for the leaders in the four workplaces where we gathered data to be interviewed about their approach to leadership. The interviews provided useful information and valuable insight to supplement the workplace interactions that had been collected. The extract below, taken from the report back to the workplace, demonstrates how the interview illuminated the way leadership was enacted in this workplace (all names are pseudonyms).

It is readily apparent that Paul is a highly motivated and accomplished individual. Everything he does, he does to his fullest extent, whether it is in his professional work or his numerous community activities. As he himself remarks, he is a ‘perfectionist’ who is always keen to take on tough challenges and prove himself. He is also always thinking about how else he might have done things, as well as constantly developing ‘contingencies’ in case things do not work out as planned. He is a firm believer in leading by example. In this extract, we see Paul providing such an example. Although overtly he is encouraging Brendan to help Imogen develop within her role, he is at the same time helping Brendan develop within his.

- Paul: So what I'm saying is that it's up to yourself and to Anna—I've said this to Anna as well—to manage Imogen, okay? And any escalating issues that come about with regards to her performance, then they come to me and that's when I'll step in and talk with her. I believe—and you'll agree I'm sure—that she has absolutely got the capability, but it's just a bit more guidance that she needs.
- Brendan: Oh she has. Yeah.
- Paul: And I think through this learning curve—although it's been a pretty hard one for you to take and a hard one for her to understand—
- Brendan: Well it could have been a lot worse, I think.
- Paul: Yeah. I think it'll come out on the other side then a lot better for the experience, yeah.
- Brendan: Yeah and she'll appreciate the time that we've put in.

In a more recent project, workplace data has been supplemented not only with interviews but also with a complex range of other data. In following professional migrants who were completing a communication course at Victoria University of Wellington, we recorded workplace interactions as they undertook internships in local businesses and government organizations. We also recorded interviews with them and their workplace mentors and a range of course-based data. This included role-plays and their views on how they performed in these role-plays, as well as oral presentations and group discussions (see Riddiford and Joe 2010: 197). The resulting rich database of material enabled the course coordinators to track the development of the skilled migrants' sociopragmatic performance over the duration of the course, from the classroom to the workplace. It also meant that the participants were encouraged to think about their own communication choices and see how they could adapt these in order to communicate more efficiently.

3.5 Summary

Engagement with the workplaces and the volunteers who have worked with us has had a number of very positive outcomes. The participants' responses in workshops and in interviews have provided valuable insights

as well as validation for our interpretation of their data. They have also highlighted many future areas of potentially fruitful research. We have also been able to give something back to the workplaces, through both workshops and written reports, validating and providing positive reinforcement for the effective and positive communication that we have observed in these workplaces. And using the CED model and working with the skilled migrant workers who participated in our research, we have helped people reflect on and develop strategies for improving their communication skills.

4 Engagement with the Wider Community

4.1 Exploring Further Outlets for Reporting Research Findings

Workshops and reports proved to be useful ways of providing many workplaces with information on our research results. We have also written summaries of our research for workplace and industry newsletters. For instance, after collecting data in a hospital ward, feedback was provided to the wider community of nurses through the health board staff newsletter (Holmes and Major 2002a, b) and also to the community of nurses in New Zealand through local nursing journals (Holmes and Major 2003a, b).

Local business magazines have also provided a useful outlet for disseminating our research to the wider community, with a diverse range of topics being covered, from the importance of humour and small talk (Holmes 1998a, b, e, 1999a; Holmes and Stubbe 1998), to what makes a successful meeting (Holmes 1998f), the use of narrative at work (Holmes 2007) and features of the talk of managers (Holmes 1998d, 2003). These short articles enable us to highlight for the wider community a few important issues that the workplaces have found useful, succinctly summarizing our main findings in a way that is accessible to a non-specialist audience. People have been very interested, for instance, in what we have discovered about the importance of social talk in New Zealand workplaces. Social talk is often undervalued because it is not 'real work talk'.

However, our research demonstrated its importance as a means of building relationships with colleagues and helping people feel good about work within a New Zealand business context, making this a valuable point to bring to people's attention within the wider community.

The role of humour is also often undervalued in the workplace. The following extract is taken from an article published in *Employment Today*, New Zealand's leading independent human resources and employment law magazine:

In an item headed 'No jokes please, we're British' the Independent recently reported that, according to a survey of British workplaces, most UK directors took a dim view of humour in meetings, and 'a staggering 39%' believed jokes had no place at all at work! It is good to be able to report that New Zealand managers appear to be much more sensible on the issue of humour at work. Indeed, research in New Zealand workplaces shows that humour is a very important component in maintaining good relationships at work. When faced with a criticism for leaving the wrong date on a memo, for instance, instead of snapping back, one manager responded 'ah well I find it hard being perfect at everything'. His good-humoured ironic quip was well received and restored good relations. (Holmes 1999a)

Presenting our research to a wider audience within the business community through such resources as *Employment Today*, rather than publishing the results only in academic publications, is an important step in disseminating our research to a wider audience, with the opportunity for practical applications. This magazine is targeted at 'HR professionals in private and public sector organizations, CEOs, managers, supervisors and team leaders'.¹ As noted above in the overview of the project, an aim that was identified at the outset was to provide 'relevant input to human resource development programmes'. Making our material and insights available to workplace practitioners through these types of publications allows them to be accessible to people who are dealing with practical issues in managing people and who are implementing HR programmes.

The LWP research has also been discussed in popular radio programmes, and the media have generally been quick to pick up on some

¹ See <http://www.employmenttoday.co.nz/databases/modus/pages/et-contact-details>.

of the research findings and publicize these more widely within the local community. The Victoria University of Wellington magazine, *Victorious*, which is sent out to alumni, has highlighted our research on a number of occasions and local papers have regularly picked up on media releases made by the university on current research. For example, a topic that caught the media's attention after a university media release at the beginning of 2012 was the distinctiveness of workplace talk in New Zealand. A report in the local Wellington paper, *The Dominion Post*, highlighted the informality of language in New Zealand workplaces, as well as the influence of Maori norms of interaction.

Professor Janet Holmes has also been involved along with a number of other linguists in writing a column for *The Dominion Post*. The column has covered a range of language-related topics, and Professor Holmes' contributions have often included a focus on workplace communication. In one column for instance, she explores the influence of workplace culture on leadership style. This is one of the most important factors in accounting for differences in the way that leadership is enacted, and has certainly been found to be more influential than gender. These short articles are also available on the *stuff.co.nz* website and some have been collected and published in a book (Bauer et al. 2011).²

Publicity for the LWP through local magazines, newspapers, radio and websites has also led to presentations to a number of groups and organizations, such as Rotary, the Department of Labour (now Ministry of Business, Innovation and Employment), Relationship Services, and the New Zealand Employment Court judges.

We have also produced a range of brochures on different topics, such as small talk, humour and giving directives, which we have taken to workplaces and sent out to people who have requested information about our project. Nowadays these brochures can be readily accessed through our website,³ which provides an excellent platform for presenting information on a range of the topics we have researched. Short summaries of many of these topics are posted on this website. The example below, for instance, is taken from the entry on narrative.

² For the *stuff.co.nz* website, see <http://www.stuff.co.nz>.

³ See <http://www.victoria.ac.nz/lwp>.

Workplace anecdotes are a special type of narrative. They are brief digressions from core-business that serve social purposes, such as humanising the narrator and/or downplaying positions of authority.

e.g. Opening of weekly team meeting in a large commercial organisation

Peg: *well welcome back Clara*

Clara: *thank you Peg*

Peg: *we really missed you*

Clara: *oh thank you Peg*

All: *[laugh]//[drawl]:oh*

Peg: */(he was) really awful to us you don't know what he made us do [laughs]*

All: *[laugh]*

Clara: *the //power went\to his head did it?*

Sandy: */(did not)*

Peg: *oh it was just terrible and he kept going up and sitting in your desk [laughs] [...]*

The content of the story may not be particularly significant to the listeners, but the function of the story in creating a particular image of the narrator means that anecdotes are an important part of workplace interactions.⁴

The project's website makes such summaries accessible to other researchers, students or anyone with an interest in our project. It also provides a bibliography of all our research so that anyone interested in a particular topic can easily find further references to papers which provide more detail on that subject.

A further aim in setting up the website was to provide easy access to other resources. Available here, for example, are a number of occasional papers on a range of topics, such as data collection issues across diverse workplace contexts, as well as literature reviews and research reports related to specific projects. A practical guide to the transcription conventions we use is also available here.⁵ This enables us to share practical information on data collection and processing with researchers from around

⁴ See <http://www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace/research/narrative>.

⁵ See <http://www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace/publications/occasional-papers>.

the world and there are now several projects using our methodology (see, for example, Richards 2006; Mullany 2007; Handford 2010; Angouri and Bargiela-Chiappini 2011; Ladegaard 2011).

4.2 Providing Input for Education

Since the beginning of the LWP, there has been a focus on making the results accessible to English for speakers of other languages (ESOL) teachers through both local and international publications (for example, Holmes 1999b, 2000b, 2001, 2002, 2009; Newton 2004; Malthus et al. 2005; Holmes and Riddiford 2011). The LWP website also provides details of other resources that have been produced by members of both the core and wider team to feed into teaching and education programmes.

Directly linked to our aim of providing relevant input to human resource development programmes, a resource kit was produced in 2002 (Stubbe and Brown 2002). This is a flexible training resource for exploring the communication strategies used in effective multicultural workplace teams and learning about the discourse features of spoken English in typical New Zealand factories. The kit includes a video and handbook for workplace trainers, English language teachers, human resources professionals and others involved in the development of workplace language, literacy and communication skills. The materials are based on actual footage of workplace interaction and interviews with factory staff.

Another teaching resource which has been produced based on the LWP data is a textbook designed for business/workplace ESOL classes or communication training courses (Riddiford and Newton 2010). Each unit in the book is based around recordings of workplace interactions gathered by the LWP team as they naturally occurred in a range of professional workplaces. The units each contain a range of activities to encourage reflection, discussion, analysis and communication practice focused on particular kinds of interaction and speech functions that are difficult to manage interculturally, such as requesting, refusing, disagreeing, complaining and apologizing.

This textbook was produced from materials developed for a course at Victoria University of Wellington preparing skilled migrants for New Zealand workplaces (The Workplace Communication Programme for

Skilled Migrants). This course began in 2005 as the result of the Tertiary Education Commission (the official body through which the New Zealand government funds tertiary education) making funds available for the development of courses to get unemployed and underemployed skilled migrants into skilled occupations. The course focuses on helping these skilled migrants 'to develop their sociopragmatic competence, in order to increase their chances of gaining employment in their chosen professions' (Riddiford and Joe 2010: 195; see also Riddiford and Joe 2005).

The data that had been collected over the years by the LWP, together with the analyses that had been conducted, highlight a number of important aspects of communication in New Zealand workplaces that at times prove difficult for migrants. The teachers working on the skilled migrant course wanted to employ the insights gained to help migrants find work. Analysing examples adapted from the LWP corpus enables students to better understand the pragmatics of New Zealand English, including the importance of small talk, and the challenges of making requests and refusals appropriately in a new sociocultural context.

In the materials on refusals, for example, students are given a scenario based on a real interaction. They are asked to consider aspects of the context and then to role-play the refusal, having been given the first couple of utterances from a modified transcript to start them off. Next they compare their role-play with the modified transcript of the real interaction. The following example shows a modified transcript involving a refusal.

1. Nicola: Well ... um ... the thing is that the minister needs to be
2. briefed remember we talked about that?
3. Claire: Yeah.
4. Nicola: And that you did the original brief and Tom's not
5. wanting me to do the brief ... because it's not our work.
6. Claire: Oh, and you want to bring it over here?
7. Nicola: Yeah, and so we were wondering if you could do the brief
8. because ... we're not going to ... and because you've got the
9. first one.
10. Claire: [drawls] Oh.
11. Nicola: And we were just hoping you could whittle down what
12. you wrote last year and just ...

13. Claire: The problem I've got is that, um, that, um ... Joseph—well
14. not Joseph ... The thing is that the managers haven't
15. decided ... where this work's going to fall ... so ... I won't be
16. able to do it officially ... informally yeah, but the thing is
17. that, um, this is the problem at the moment, the managers,
18. er, haven't decided where the work's going to fall.
19. Nicola: Mm, yeah.

Students are asked to complete a number of tasks related to the transcript and its associated recording, which has been rerecorded with actors to match the modified transcript. Engagement with the data in this way raises the students' awareness of the strategies which native speakers use; the feedback we have received indicates that students find these resources very helpful.

The course at Victoria University of Wellington is aimed at professional migrants. There are also many migrants and refugees in New Zealand who do not have professional qualifications and who therefore tend to find work in blue-collar contexts. Two industries which have been identified as ones where migrants, and in particular refugees, are currently seeking employment in New Zealand are the construction industry and eldercare residences. The construction industry is employing many migrants to aid with the rebuild of Christchurch after the earthquakes in 2010 and 2011, while New Zealand eldercare facilities employ a high number of migrants working as caregivers.

These industries were not areas where the LWP had previously collected data, so new workplaces were identified for data collection with a view to directly feeding the data into the development of teaching materials. Data was collected by builders on two building sites as they worked on constructing residential housing. The eldercare research focused on interactions between caregivers and residents, this time at three different sites. Both sets of data were then examined for the language and ways of speaking used by the interactants as they went about everyday tasks. Dialogues were extracted which showed, for instance, the foreman

directing his apprentice to complete a task, or a caregiver engaging in small talk with a resident. These extracts were modified and recorded with 'new' voices to provide audio materials for use in teaching. Migrant workers in the construction industry may not have high levels of literacy in English, so the development of teaching materials required a quite different approach from that employed in producing materials for the skilled migrant course. Audio is also used in the professional materials, but the construction site materials are picture based and include a great deal of repetition for practice purposes.

The construction site materials were trialled with a local class taught through an organization called MCLaSS (Multicultural Learning and Support Services). This non-profit incorporated society offers free education and support for adult refugees and migrants in the Wellington region. Nicky Riddiford used the materials with the class while their normal teacher sat in. The students had a low level of proficiency in English and also low literacy skills. Although they were not aiming to work in the building industry, they were all doing work experience in different trades. This trialling allowed Nicky Riddiford to gauge whether the materials were providing sufficient input for the students; valuable feedback from the teacher also provided suggestions which she incorporated in the final version.

In 2013, these completed materials were added to our website for free downloading. This provides easy access for anyone who is involved in teaching and working with migrants seeking employment in these industries. Feedback received to date suggests that teachers are finding it very useful to have access to New Zealand-based resources.

The practical teaching resources and input that we have been able to provide has enabled us to fulfil a primary goal of our research, namely that it should feed into human resource development programmes. We have extended this by specifically targeting a human resource issue which was identified by the government, that is, the need to provide resources and support for migrants seeking work in New Zealand and for practitioners working with them.

4.3 Advice and Development of Materials for Government Agencies

LWP research has also been used to inform advice and materials provided to government agencies to aid them in producing resources for migrants and employers. The LWP were approached by the Settlement Division of the (then) Department of Labour (now the Ministry of Business, Innovation and Employment), who had become aware of our project through the skilled migrants course and who felt that our work would 'add value' to the work that they do. Their goal is to 'maximize the potential of an increasingly mobile global workforce' (Ministry of Business, Innovation and Employment 2013: 11). This collaboration has led to the production of two highly successful resource kits: one for migrants to New Zealand and one for employers of migrants.⁶ These resources are also being tailored to different industries and to migrants from different places. For example, a set of resources aimed at the construction industry has been released, along with one giving advice and support for migrants from the Pacific Islands.

This collaboration has also led to the development of *Work Talk*, a website designed to improve communication between employers and new migrant employees.⁷ The site provides five scenarios where migrants interact with more senior workers who are New Zealanders and who employ New Zealand ways of speaking and interacting. There are two ways of navigating the website, one for employers who may have migrant employees and one for the migrant employees themselves. For example, one scenario shows a migrant waiting for the lift. He is joined by a manager, who greets him and tries to engage in small talk, but the migrant gives only minimal responses and keeps his head down. Many new migrants find this situation challenging: they come from a culture where they are not used to interacting on an informal basis with their boss. Migrants are asked to choose what they would do if they were the employee and employers are asked what they would have done as the manager. Feedback to migrants highlights the communication issues from the point of view of the manager; the migrant

⁶ See <http://www.ssnz.govt.nz/living-in-new-zealand/information-resources/index.asp>.

⁷ See <http://worktalk.immigration.govt.nz>.

is expected to make eye contact and to respond more fully to the manager. He could also ask the manager how his day is going or say something about the weather. Feedback to the manager highlights the fact that the migrant comes from a culture where workers do not interact informally with their boss. There is then a 'best choice video' which shows the migrant looking up and responding positively to the manager's attempts to engage him in conversation.

A major aim of the website is to raise awareness in both groups of different ways and patterns of interacting. The migrants represented on the website have a range of backgrounds and the focus is on language and norms of interaction that may cause miscommunication on both sides.

4.4 Summary

Short popular articles summarizing our research make it accessible to a wide audience. This has led to our research being picked up by local media, which has in turn led to requests for presentations by a range of public and private groups and organizations. Using our insights to highlight a number of communication issues in these types of forums means that other workplaces we have not researched can also benefit. The workplaces who have worked with us are part of the New Zealand business community, so the insights gained from working with them have implications and applications for the wider community. The LWP website has also provided a valuable resource not only for academics, teachers, students and workplaces, but also for members of the general public who may have an interest in our work.

Input into education through the use of our data in the Skilled Migrants' Course and the subsequent publication of a textbook drawing on this data, extends the practical application of our research findings. Targeting resources aimed at migrants wanting to work in specific industries takes this a step further and our ongoing collaboration with a government ministry allows us to give direct input into practical resources produced by another organization. Engaging with a government unit whose specific aim is to assist new migrants enables us to offer advice and materials which can help both employers and employees to raise their

awareness and understanding of the issues involved. The websites also provide an opportunity for people outside New Zealand who are considering coming here to access relevant resources and information based on authentic interaction.

5 Conclusion

This chapter has outlined a range of approaches that the LWP team has taken in engaging with the public. The LWP was conceived from both an academic and an applied viewpoint from the outset and this has had important implications for all aspects of the research, from the initial and ongoing design and aims, to the methodology, to the ways that the results of analysis have been presented and utilized. Community-based research such as this has the underlying assumption 'that research is not simply an intellectual act, but that it is also a practical act that can have practical implications and applications' (Czaykowska-Higgins 2009: 26). Importantly for the LWP team, sharing insights and research results began with the workplaces where data had been collected. The partnerships which developed with workplaces provided valuable input into our own research goals, as well as broadening understanding of workplace communication on both sides. Researching *with* rather than *on* participants produced a much richer database and enabled us to give something back to the people who worked with us.

Being able to share these insights with the wider business community through a number of mediums, rather than publishing them only in academic publications, highlighted the practical importance of the research findings, which in turn further validated the input from workplaces. Industry newsletters and magazines, local media and seminars to organizations such as Rotary, all provided outlets for us to reach out to the wider community. The project's website has also provided a platform for making available a wide range of resources of different kinds. Educational materials are one kind of resource that the website offers. Input into education is another important way in which the research that the project undertakes can benefit the wider community. Training materials which draw on authentic data are rare, so teachers have been very keen to access these and feedback has shown that they find them very useful.

A final way that we have engaged with the wider community is our work with the Settlement Division of the Ministry of Business, Innovation and Development. This collaboration underlies a recognition of the crucial role that communication in the workplace plays in achieving this government ministry's goal of helping organizations be productive. Drawing on the research of the LWP, they aim to provide the best resources they can to aid migrant employees and their employers.

Engagement with a range of communities has enriched our research in many ways. Insights from the research participants have enhanced our analyses and the partnerships forged have opened new avenues for research as we have been able to respond to the needs of the workplaces and the wider business community. We believe the huge benefits in adopting a community-based approach in research are very evident and that the implications and practical applications that have arisen from the LWP's work will continue to benefit a wide range of people.

References

- Angouri, Jo, and Francesca Bargiela-Chiappini. 2011. 'So what problems bother you and you are not speeding up your work?': Problem Solving talk at work. *Discourse & Communication* 5(3): 209–229.
- Bauer, Laurie, Dianne Bardsley, Janet Holmes, and Paul Warren. 2011. *Q and Eh? Questions and Answers on Language with a Kiwi Twist*. Auckland: Random House.
- Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, Ben Rampton, and Kay Richardson. 1992. *Researching Language: Issues of Power and Method*. London: Routledge.
- Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous communities. *Language Documentation & Conservation* 13: 15–50.
- Hammond, Sue Annis. 1996. *The Thin Book of Appreciative Enquiry*. Plano, TX: Thin Book Publishing.
- Handford, Michael. 2010. *The Language of Business Meetings*. Cambridge: Cambridge University Press.
- Holmes, Janet. 1998a. Don't under-rate small talk. *NZ Business* April: 52.
- . 1998b. Humour gets things humming. *NZ Business* October: 60.

- . 1998c. No joking matter! The functions of humour in the workplace. *Proceedings of the Australian Linguistics Society Conference*. Brisbane University of Queensland. <http://www.als.asn.au/proceedings/als1998.html> (accessed 8 August 2015).
- . 1998d. The power of talk. *Management* May: 56–58.
- . 1998e. Small talk. *NZ Business* April: 52.
- . 1998f. What's in a successful meeting? *New Zealand Business* 19.
- . 1999a. Humour makes work go well. *Employment Today* July: 10–11.
- . 1999b. Managing social talk at work: What does the NESB worker need to know? *TESOLANZ Journal* 7: 7–19.
- . 2000a. Doing collegiality and keeping control at work: Small talk in government departments. In *Small Talk*, ed. Justine Coupland, 32–61. London: Longman.
- . 2000b. Talking English from 9 to 5: Challenges for ESL learners at work. *International Journal of Applied Linguistics* 10(1): 125–140.
- . 2001. Implications and applications of research on workplace communication: A research report. *New Zealand Studies in Applied Linguistics* 7: 89–98.
- . 2002. Workplace communication and ESOL teaching. *CLANZ* 3: 8–10.
- . 2003. Women's talk at the top. *Boardroom: The Journal of the Institute of Directors* May: 1.
- . 2007. Telling tales at work. *New Zealand Management* May: 36–38.
- . 2009. Is sex relevant in the ESL classroom? In *Best of Language Issues*, eds. Rakesh Bhanot, and Eva Illes, 272–278. London: London South Bank University Press.
- Holmes, Janet, and Rose Fillary. 2000. Handling small talk at work: Challenges for workers with intellectual disabilities. *International Journal of Disability, Development and Education* 47(3): 273–291.
- Holmes, Janet, Rose Fillary, Marianne McLeod, and Maria Stubbe. 2000. Developing skills for successful social interaction in the workplace. *New Zealand Journal of Disability Studies* 7: 70–86.
- Holmes, Janet, and George Major. 2002a. 'It's getting a bit desperate isn't it!' Communication on the Ward. *Staff Matters, Staff Newsletter of Capital & Coast District Health Boards* June 41: 8.
- . 2002b. 'Just flex your arm eh'. Communication on the Ward. *Staff Matters, Staff Newsletter of Capital & Coast District Health Boards* July 42: 13.
- . 2003a. Nurses communicating on the ward: The human face of hospitals. *Kai Tiaki: Nursing New Zealand* 8(11): 14–16.

- . 2003b. Talking to patients: The complexity of communication on the ward. *Vision—A Journal of Nursing* 11(17): 4–9.
- Holmes, Janet, Meredith Marra, and Bernadette Vine. 2011. *Leadership, Discourse and Ethnicity*. Oxford: Oxford University Press.
- Holmes, Janet, and Nicky Riddiford. 2011. From classroom to workplace: Tracking socio-pragmatic development. *ELT Journal* 65(4): 376–386.
- Holmes, Janet and Maria Stubbe. 1998. Small talk, business talk—Oiling the wheels of business. *People & Performance* June: 28–31.
- Holmes, Janet, and Maria Stubbe. 2003. *Power and Politeness in the Workplace: A Sociolinguistic Analysis of Talk at Work*. London: Longman.
- Jones, Deborah, and Maria Stubbe. 2004. Communication and the Reflective Practitioner: A shared perspective from sociolinguistics and organisational communication. *International Journal of Applied Linguistics* 14(2): 185–211.
- Ladegaard, Hans J. 2011. ‘Doing power’ at work: Responding to male and female management styles in a global business corporation. *Journal of Pragmatics* 43(1): 4–19.
- Major, George, and Janet Holmes. 2002. Capital Coast Health Nurse–Patient Communication Study Methodology Notes. *Language in the Workplace Occasional Papers* 8. Available at <http://www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace/publications/occasional-papers> (accessed 8 August 2015).
- Malthus, Caroline, Janet Holmes, and George Major. 2005. Completing the circle: Research-based classroom practice with EAL nursing students. *New Zealand Studies in Applied Linguistics* 11(1): 65–89.
- McCallum, Judi. 2013. Working in an Eldercare Facility: An ESOL Resource—Unit 1. Language in the Workplace Project. Available at <http://www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace/resources/teaching-and-learning-resources> (accessed 8 August 2015).
- Ministry of Business, Innovation and Employment. 2013. *Briefing to Incoming Minister of Immigration*. <http://www.mbie.govt.nz/pdf-library/about-us/bims/MBIE%20Immigration%20BIM.pdf> (accessed 9 August 2015).
- Mullany, Louise. 2007. *Gendered Discourse in Professional Communication*. Basingstoke, New York: Palgrave Macmillan.
- Newton, Jonathan. 2004. Face-threatening talk on the factory floor: Using authentic workplace interactions in language teaching. *Prospect* 19(1): 47–64.
- Richards, Keith. 2006. *Language and Professional Identity: Aspects of Collaborative Interaction*. Basingstoke: Palgrave Macmillan.
- Riddiford, Nicky, and Jonathan Newton. 2010. *Workplace Talk in Action—An ESOL Resource*. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

- Riddiford, Nicky. 2013a. Working on a building site: An ESOL resource—Unit 1. Language in the Workplace Project. Available at <http://www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace/resources/teaching-and-learning-resources> (accessed 8 August 2015).
- . 2013b. Working on a building site: An ESOL resource—Unit 2. Language in the Workplace Project. Available at <http://www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace/resources/teaching-and-learning-resources> (accessed 8 August 2015).
- . 2013c. Working on a building site: An ESOL resource—Unit 3. Language in the Workplace Project. Available at <http://www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace/resources/teaching-and-learning-resources> (accessed 8 August 2015).
- Riddiford, Nicky, and Angela Joe. 2005. Using authentic data in a workplace communication programme. *New Zealand Studies in Applied Linguistics* 11: 103–110.
- . 2010. Tracking the development of sociopragmatic skills. *TESOL Quarterly* 44(1): 195–205.
- Stubbe, Maria. 1998. Researching language in the workplace: A participatory model. *Proceedings of the Australian Linguistics Society Conference. Brisbane University of Queensland*. July 1998. <http://www.als.asn.au/proceedings/als1998/stubb266.html> (accessed 8 August 2015).
- . 2001. From office to production line: Collecting data for the Wellington Language in the Workplace Project. *Language in the Workplace Occasional Papers* 2. Available from <http://www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace/publications/occasional-papers> (accessed 8 August 2015).
- Stubbe, Maria, and T. Pascal Brown. 2002. *Talk that works: Communication in successful factory teams: A training resource kit*. School of Linguistics and Applied Language Studies. New Zealand: Victoria University of Wellington.
- Stubbe, Maria, and Bernadette Vine. 1998. Swings and roundabouts: Getting things done at work. *Tē Reo* 41: 182–188.

Index

A

- academic discourse, 16, 295
- Academic Writing in English (AWE),
 - 14, 211, 216, 223–7, 232,
 - 233, 237
- accommodation, 148, 169, 244,
 - 254
- adjacency pairs, 303, 305, 308, 309
- Africa, 30, 251, 261
- analogue, 109, 184, 231–2
- annotation, 4, 9, 11, 17, 26, 39, 40,
 - 42, 48n21, 62, 138,
 - 187n12
- AntConc, 37, 37n14, 41, 58, 194,
 - 194n17, 205
- applied linguistics, 243–82, 311
- apps, 4, 8, 10, 14, 211–37
- archival recordings, 104, 110, 129,
 - 146
- AspMAP, 111
- Audacity, 109n8
- audiences
 - creative industries, 40
 - employers, 340
 - heritage sector/industries, 26, 40,
 - 178
 - human resource(s), 333
 - lifelong learners, 89
 - migrants, 10, 40, 53, 62, 341
 - museum, 10, 13, 40, 144
- audio, 2, 9, 15, 58, 61, 73, 90, 103,
 - 109, 110, 117, 122, 134,
 - 135, 137–41, 144–8, 150,
 - 165, 181, 184, 185, 187,
 - 190–3, 195, 196, 198, 200,
 - 267, 268, 276, 281, 286,
 - 295, 323, 326, 339
- audio files, 11, 29, 59, 110, 117,
 - 140, 163, 192, 198, 277,
 - 278

B

- Bank of English, 72
- BBC Voices*, 75, 100, 100n1, 121n13, 126
- Belgium
 - Belgian, 5
 - Flanders, 253
- best practice, 2, 9, 12, 14, 15, 17, 26, 40–2, 61, 62, 128, 172, 180
- birth dates, 104, 106, 184
- blog, 32, 49, 39n15, 55n30, 77, 78, 81, 83, 87, 91, 93, 123, 278, 281
- books (non-specialized audiences), 147
- British Academic Spoken English (BASE), 294
- British Academic Writing in English (BAWE), 223, 224
- British Component of the International Corpus of English (ICE-GB), 215, 218, 227, 232, 234, 235
- British National Corpus (BNC), 72, 228

C

- Canada
 - Labrador, 11, 101, 102, 111, 125
 - Newfoundland, 11, 101, 102, 111, 125
- cartographic tool, 255
- cassette tape, 107, 135
- CD (audio), 11, 12, 59, 163, 165, 206
- chat room, 32
- cloze test/cloze exercise, 229, 254

- code-switching, 250
- Collins Birmingham University International Language Database (COBUILD), 1
- comma-separated values (CSV), 127
- Communication Evaluation and Development (CED), 329–30, 332
- communities of practice, 307, 313
- community-based approach, 322, 325, 343
- community-based research, 144, 342
- community development, 160
- community identity, 161, 172, 173, 273
- concealed judgement task, 254
- concordance/concordancing, 73, 141, 220, 234, 278, 314
- concordancer, 73
- concordancing programs, 141, 220, 234, 314
- continuing professional development (CPD), 9, 15–17, 195, 200, 205, 311
- conversation, 12, 16, 31, 32, 35, 58, 59, 160, 162, 182, 183, 186, 188, 212, 274, 292, 303, 341
- conversation analysis, 16
- copyright clearance, 7, 91, 128
- corpus linguistics (CL), 14, 16, 17, 90, 214, 215, 291–4, 296–303, 305, 307–10, 313, 314
- corpus linguistics and conversation analysis (CLCA), 16, 291–3, 296, 310–14

- Corpus of Irish English
 Correspondence
 (CORIECOR), 26, 27,
 27n3, 29, 32, 34–42,
 40n17, 44, 52, 53n28,
 55–9, 61
- Corpus of Modern Scottish Writing
 (CMSW), 11, 70, 75–8, 85,
 87, 89–92, 94
- Corpus Presenter 14, 36
- correspondence corpora, 26, 27, 30,
 39, 42, 52, 61, 62
- courses
 e-textbook, 215, 225, 226
 exercises, 30, 88, 205, 212, 214,
 216–36, 277
 gamification, 222
 online, 132
 self-study, 223
- critical discourse analysis (CDA)
 centring, 313
 othering, 306, 313
- crowdsourcing, 80
- D**
- database(s), 1, 10, 12, 14, 25–62,
 80, 82, 90, 110, 111,
 111n9, 116, 119, 120, 126,
 127, 129, 135–7, 139–42,
 144, 220, 231, 233, 234,
 243–62, 322–5, 331, 342
- database tables, 139
- data extraction, 108, 110
- data management, 4, 11, 133–52
- data protection, 28
- debt incurred, principle of, 2, 12,
 134, 272
- diachronic, 35, 38, 57
- Diachronic Electronic Corpus of
 Tyneside English (DECTE),
 4, 7, 12, 13, 177–207, 286
- Dialect Atlas of Newfoundland and
 Labrador, 11, 99–129
- dialects
 acquisition, 75
 areas, 59, 101
 atlas, 5, 99, 100, 102, 103, 119, 129
 diversity, 135, 147
 geography, 100
 input, 106
 lexicon, 107, 162, 229
 mapping, 99, 101, 102
 questionnaire, 107
 variation, 15
- diaspora, 10, 25–26, 42, 108
- Dictionary of the Scots Language
 (DSL), 80, 82
- Digital Dictionary of American
 Regional English (Digital
 DARE), 109
- digital humanities, 39, 86n10
- Digitising Experiences of Migration
 (DEM), 26, 39–52
- Digitizing Immigrant Letters (DIL), 50
- discourse-pragmatic features
 directives, 327, 334
 discourse marker *like*, 34
 discourse markers, 34, 38, 272,
 294, 301, 310
 hedges, 272
 humour, 327, 328, 332–4
 intensifiers, 281, 285
 quotatives, 270, 272, 285
 small-talk, 327, 332, 334, 337,
 339, 340
- Diversity in Dutch DP Design
 (DIDDD), 14, 244, 252, 255

- documentaries, 58, 80, 134, 144–6
- Documenting Ireland: Parliament,
People and Migration
(DIPPAM), 25–7, 29, 36,
39, 40, 40n17, 50, 61, 82
- Dynamic Syntactic Atlas of the
Dutch Dialects
(DynaSAND), 14, 243,
244n2, 252–6, 258
- E**
- education materials, 134, 150
- ELAN, 187, 187n12
- email, 77, 84, 90, 93
- encoding, 27, 41–3, 45, 46, 52, 62
- England
- Derby, 183
 - Essex
 - Havering, 272, 273
 - Romford, 273
 - London
 - Hackney, 273
 - Islington, 273
 - Wood Green, 273
 - North East England
 - County Durham, 185
 - Gateshead, 184, 185
 - Middlesbrough, 185
 - Newcastle, 179n4, 185
 - Redcar & Cleveland, 185
 - Teesside, 185
 - Tyneside, 185
 - Wearside, 185
- English as a Foreign/Second
Language (EFL/ESL). *See*
English for speakers of other
languages (ESOL)
- English for speakers of other
languages (ESOL), 336
- English Language Teaching, 211,
265–88
- English Language Teaching
Resources Archive, 269–78
- English Spelling and Punctuation
(ESP), 14, 211, 216, 223,
227–30, 232, 233, 235, 236
- Enhanced British Parliamentary
Papers on Ireland (EPPI),
27
- Enhanced Repository for Language
and Literature Researchers
(ENROLLER), 50, 62, 92,
189, 189n15
- ethics/ethical, 15, 101, 152, 192,
198, 272–8, 277, 325
- ethnicity, 36, 246, 249
- ethnomethodology, 294, 303
- events
 - Great Irish Famine, 40
 - Scottish devolution, 72
 - Vernacular revival (Scotland), 76
- exhibitions, 6, 11, 37, 52–5, 55n30,
77–81, 101, 102, 112, 134,
146
- explanatory clarification, 32
- Extensible Mark-up Language
(XML), 4, 36, 48n20, 49,
50n26, 62, 180, 180n6,
181, 184, 185, 188–95,
197–202, 206, 207
- Extensible Stylesheet Language
Transformations (XSLT),
50, 50n26
- F**
- Facebook, 80, 81, 87, 88, 91, 93, 94,
123
- false starts, 32, 188

feedback, 18, 87–9, 125, 221, 226,
 231, 233, 234, 236, 281,
 322, 324, 326–9, 332,
 338–41
 fixed-format elicitation, 12, 162
 focus group, 88, 89
 forensic, 261
 forensic linguistics, 292
 formality, 301
 France, 253, 254, 281
 French Flanders, 253

G

Gabmap, 101
 games, 102, 121–4, 205, 222,
 303–7, 309
 gender, 35, 36, 88, 105, 114, 117,
 167, 168, 196, 245, 253,
 272, 327, 334
 generational shift, 166
 geographical, 35, 36, 46, 49, 76, 80,
 106n5, 147, 246, 249, 261,
 266
 Glow (Scottish Schools National
 Intranet), 80, 81, 89
 Goeman, Taldeman, van Reenen
 Project (GTRP), 14, 243–4,
 244n2, 252–5
 Goldvarb, 37
 Google Analytics, 91, 281
 Google Maps, 73, 74, 127
 grammar, 10, 12, 14, 15, 38, 77,
 104, 113, 121, 122, 200,
 213–16, 218, 222, 226,
 232–4, 237, 266
 grammatical complexity, 235
 graphicalization, 138
 graphological features, 62
 graphology, 42

H

handwriting, 43, 77
 Helsinki Corpus of Older Scots, 76
 higher education (HE), 2, 3, 8, 10,
 16, 37, 177–80, 207, 287
 Historical Thesaurus of English, 82,
 112
 history (social), 41
 HyperText Mark-up Language
 (HTML), 50n25, 120n12,
 124

I

immigrant languages, 143
 immigrant(s), 30, 38, 53, 57, 147,
 249, 268
 impact, 2, 4, 5, 9, 10, 12–14, 25–62,
 70, 84, 92, 93, 102, 124–5,
 164, 178, 179, 207, 214,
 248, 272, 288, 330
 informal, 32, 34, 78, 88, 91, 104,
 186, 213, 295, 323, 340
 informality, 34, 334
 informants
 anonymity, 6, 273, 274
 anonymization, 199, 275
 anonymization protocol, 7, 15, 277
 anonymizing, 198, 277
 confidentiality, 273, 274
 consent, 6, 276
 permissions forms/permission
 tracking, 90
 pseudonym(s), 270, 328
 substitution, 192, 198
 ingroup/in-group, 246, 275, 328
 interaction, 16, 76, 139, 140, 232,
 286, 291, 293–6, 301–3,
 308, 310, 313, 323, 324,
 327, 334, 336, 337, 341, 342

- interactive Grammar of English (iGE), 14, 211, 216–23, 226, 232, 233, 234n6, 255
- interactive presentation, 195
- interactive website, 100
- interactivity, 121–4, 215, 225–32
- interface
- design, 218, 231
 - user, 219
 - web, 136, 192, 198, 199, 207, 219
- International Phonetic Alphabet (IPA), 106n5, 110, 117, 119, 120n12, 258
- Internet Grammar of English (IGE), 215, 215n3, 216, 218, 220, 231
- interoperability, 27, 27n1, 62, 189
- interruption, 190, 191, 194, 254, 270
- interview
- generational, 166
 - informal, 104, 186
 - one-on-one, 182
 - protocol, 56, 162, 186
 - semi-structured, 73
 - telephone, 101, 254
- Ireland, 25, 27, 28, 30, 31n7, 32, 33, 35, 45, 52–7, 59, 61, 101, 103, 106, 129, 202n19, 265
- Leinster, 36
- Irish Emigration Database (IED), 27, 29, 50
- J**
- journals/diaries, 16, 28, 134, 148, 226, 269, 278, 288, 322, 332
- judgement task, 254
- K**
- knowledge dissemination, 111
- knowledge exchange, 69, 70n1, 71, 93, 269
- knowledge mobilization, 112, 125
- knowledge transfer, 211
- L**
- L1, 261
- L2, 3, 261
- LANCHART Corpus, 136
- Language Analysis for the Determination of Origin (LADO), 10, 14, 244–52, 257, 258n11, 259n12, 260, 261
- language change, 35, 53, 76, 125, 161, 268, 269
- language contact, 56, 250
- language history from below, 29
- Language in the Workplace Project (LWP), 321–43
- language realignment, 161
- layout, 77
- legacy recordings, 102
- lexical frequency, 235
- lexical surveys, 100
- lexicon/lexical, 82, 93, 100–2, 104, 107–10, 113, 114n11, 116, 122, 127, 129, 162, 166, 229, 235, 246, 252, 254, 258, 259, 293, 296
- linguistic atlas, 101, 127n16, 166, 247, 248
- linguistic gratuity, principle of, 2, 12, 102, 134, 152, 272–3

linguistic identity, 173, 181
 Linguistic Research Digest, 269,
 278–81, 285, 288
 linguistic variation, 29, 37, 260–1
 linguistic varieties
 African American speech, 166
 African English, 261
 British, 106, 260
 Brittonic, 78
 Canadian English, 38
 Celtic, 78
 Cockney, 268
 Dutch
 Brabantish, 258
 Flemish, 258
 Limburgian, 258
 Early Modern English, 71
 Irish, 29, 38
 Irish-English, 29, 38–9
 Late Modern English, 38
 Middle English, 71
 Multicultural London English
 (MLE), 268, 276
 Newfoundland English, 116
 New Zealand English, 337
 Old English, 78
 Old Norse, 78
 Palauan English, 275
 Pictish, 78
 Scots, 71–3, 76–82, 93
 Scottish English, 71–3, 77,
 314
 Scottish Gaelic, 35
 Scottish Standard English, 71
 Spanish, 165–6
 Standard English, 38, 116, 267
 Ulster Scots, 30
 literacy, 30, 31, 253, 336, 339

M
 manuscript, 41, 77, 190
 Mapping Metaphor, 11, 70, 82–4,
 86–9, 91, 92, 94
 maps, 12, 79, 83, 99–101, 104,
 106–11, 113, 116, 117,
 119, 121, 121n13, 126,
 127, 128n17, 185n10, 255,
 256, 258, 262
 meaning questions, 125, 292
 Memorial University of
 Newfoundland Folklore and
 Language Archive
 (MUNFLA), 103, 105,
 109, 110, 128
 message board, 80, 81
 metadata, 7, 11, 17, 45–7, 49–52,
 62, 71, 73, 90, 91, 138,
 194, 197, 198
 metadata fields, 213
 Michigan Corpus of Academic
 Spoken English (MICASE),
 294
 Microcomparative Morphosyntactic
 Research tool (MIMORE),
 14, 244, 248, 252, 253n9,
 255–8, 260–2
 Microsoft Access, 111
 microvariation, 14, 243–62
 migration, 25–62, 268
 minority language, 111
 mobile learning, 211–37
 monitor corpus, 179, 180, 185–94
 morphophonology/
 morphophonological, 15
 morphosyntactic features
 articles, 255
 auxiliaries, 254

- morphosyntactic features (*cont*)
 cliticization, 254
 comparatives, 253
 complementizer, 254
 complementizer agreement, 254
 conditionals, 294
 demonstratives, 255
 diminutives, 253
 grammatical gender, 105
 modals, 33
 negation, 254
 noun phrase, 255
 object pronouns, 253
 participles, 253
 plural, 273
 possessive pronouns, 253, 255
 present tense, 105
 pronouns, 105
 pronoun systems, 105
 reciprocal pronouns, 254
 reflexives, 254
 relative clauses, 254
 semi-modals, 33, 34
 stance adverbs, 33, 34
 subject pronoun doubling, 254
 subject pronouns, 253, 254
 suffix, 252
 superlatives, 253
 verb, 252, 253
 verbal clusters, 254
 verb cluster interruption, 254
 Wh questions, 254
 will *vs.* shall, 38
- morphosyntax/morphosyntactic, 15,
 102, 105, 251, 252, 254, 255
- MP3, 117, 127
- Múin Béarla do na Leanbháin
 (MBDNL) Corpus, 34,
 34n9, 56, 57
- multicultural, 15, 267, 268, 276, 336
- multilingual/multilingualism, 15,
 249–50, 268
- museum exhibit, 134, 144, 145
- N**
- name studies, 78, 79
- narrative, 5, 28, 55, 58, 71, 183, 203,
 205, 207, 332, 334, 335
- Netherlands, 5, 248, 251, 252, 254,
 255, 257, 261
- Frisia, 252
- Newcastle Corpus of Academic
 Spoken English (NUCASE),
 16, 295–7, 308, 310–13
- Newcastle Electronic Corpus of
 Tyneside English 2
 (NECTE2), 179, 180,
 185–95, 204
- Newcastle Electronic Corpus of
 Tyneside English (NECTE),
 6, 7, 12, 13, 179, 180,
 184–5, 188–9, 192
- newsletter, 16, 73, 322, 332, 342
- newspapers, 28, 56, 101, 148, 226, 334
- New Zealand, 34, 322, 323, 329,
 332–4, 336–42
- non-standard, 33, 72, 104, 105, 122,
 183, 267
- NORMs, 107n6, 114n11
- North Carolina Language and Life
 Project (NCLLP), 134–7,
 142–7, 152
- O**
- observer's paradox, 186
- open access site, 179

- open source software, 101, 109n8, 111, 126–7
- oral history, 144, 162, 183, 186, 207
- oral history projects, 144, 183, 207
- oral narrative, 183
- Origins of New Zealand English (ONZE) Corpus, 136
- orthoepists commentary, 76
- orthographic, 73, 76, 141, 184, 190, 193
- orthographic representation, 140
- orthography, 38, 140
- outgroup/out-group, 246, 275
- overlap, 70, 82, 139, 140, 190, 246, 270, 293, 305, 307, 315
- Oxford Dictionary of National Biography (ODNB), 91
- Oxford Text Archive (OTA), 6, 8, 177
- P**
- palaeography, 80
- pamphlet, 8, 12, 163, 261, 262
- parliamentary records, 27, 28
- participation rights, 303
- part-of-speech (POS) tagging, 184, 254–6, 258
- pause(s)
- filled, 271, 272, 279
 - unfilled, 271
- pedagogy/pedagogical, 14, 211–15, 216, 220, 225, 231–4, 231–7, 294, 302
- personal letters, 29, 30, 35
- personal writing, 76
- philology, 80
- phonetic/IPA transcription, 3, 39, 71
- phonetics, 3, 39, 102–7, 109, 110, 116–20, 129, 139–41, 162, 166, 184, 185, 193, 246, 252, 254, 256, 258, 259, 266
- phonological features
- cot/caught merger, 169
 - diphthongs, 119, 166
 - fricatives, 105
 - glide shortening, 169
 - h-dropping, 187
 - ing, 105, 187
 - loss of-r, 166
 - postvocalic /l/, 106, 166
 - reversal of eI/ε, 169
 - reversal of i/I, 169
 - schwa epenthesis, 39
 - Southern Shift, 169
- Phonological Variation and Change
- in Contemporary Spoken British English (PVC), 181–6, 193
- phonology, 38, 266
- PHP Hypertext Preprocessor (PHP), 127
- picture response task, 254
- place-names, 11, 48, 78–82, 86, 88, 89, 92, 93, 146
- power, 86, 99, 236, 250, 288, 327, 328
- Praat, 139, 140
- pragmatics, 111, 233, 266, 292, 293, 337
- preference structure, 303
- prescriptive (language attitudes), 250
- public corpora, 159
- public engagement, 1–3, 7–11, 13, 14, 27, 51–2, 55, 61, 69, 70n1, 71, 82, 83, 85, 86, 89, 93, 94, 99–129, 134, 147, 177–207, 288

public-facing website, 179, 195
 public outreach, 7, 102, 112, 113,
 144–6
 public-private partnership, 12, 172
 public talks, 77, 84
 punctuation, 10, 14, 30, 211, 215,
 233, 234, 247–8

Q

quick response (QR) codes, 8, 135, 149

R

readability, 194, 219, 235
 recordings, 34, 35, 73, 75, 88,
 102–6, 108–10, 117, 129,
 134–8, 142, 144–8, 181–7,
 193, 195, 196, 198, 247,
 258, 268, 269, 275–8, 286,
 287, 295, 297, 312, 323,
 325, 326, 336, 338
 reel-to-reel tape, 103, 109, 184
 reformulation, 32
 register, 33, 55, 182, 216, 225, 227,
 233–5, 244
 relational database, 110, 111n9, 127
 relational database management
 system (MySQL), 127
 repair, 292, 303, 309
 repetitions, 32, 44, 213, 339
 representativeness, 36
 research-led teaching, 180

S

school competition, 81, 85, 86, 89
 Scotland, 71, 72, 75–8, 80, 81, 85,
 89, 91, 93, 265

Scots Words and Place-names
 (SWAP), 11, 70, 78–82,
 84–9, 91–4, 92
 Scottish Corpus of Texts and Speech
 (SCOTS), 11, 70–7, 85,
 87–92, 94, 189
 Scottish Language Dictionaries
 (SLD), 80–2, 85, 92, 93
 Scottish Place-Name Society (SPNS),
 80, 82, 85, 87, 93
 Scottish Place-name Survey, 80
 Scottish Toponymy in Transition, 89
 self-directed learning/self-learning,
 234, 237
 semantic, 82, 107, 113, 218, 266,
 293, 297, 298, 300
 semantic tagger, 297
 small group teaching (SGT), 16,
 291–316
 smartphone, 10, 14, 124, 125, 150,
 212, 215, 217, 218, 225,
 236, 237
 social media, 32, 79, 80, 82, 84, 86,
 87n11, 94, 121–4, 128n17
 social mobility, 36, 267
 social representativeness, 36
 social status, 36, 46
 social ties, 220
 social variables
 age, 245, 249, 253
 class, 245
 education, 249
 ethnicity, 36, 246, 249
 gender, 36, 245
 geographical area, 249
 geographical origin, 36, 246
 informant, 253
 literacy, 336, 339
 mobility, 36

- occupation, 196
 - race, 163, 244
 - religion, 36
 - sex, 249
 - social class, 245, 249, 253
 - Sociolinguistic Archive and Analysis Project (SLAAP), 5, 9, 11, 128, 133–52
 - sociolinguistic field methods, 150, 213
 - sociolinguistics, 11, 17, 27, 32, 36, 56, 58, 59, 71, 79, 134–7, 139, 142, 145–7, 165, 178–81, 185, 187, 243, 244, 249–50, 253, 265–88
 - socio-onomastics, 79
 - sociophonetics, 71
 - sound files, 9, 102, 109, 110, 117, 120, 127, 128, 191, 198
 - Sound Forge 8.0, 109
 - speaker-internal variation, 244
 - speakers, 6, 15, 17, 33, 34, 73, 78, 82, 90, 100–7, 110, 116, 117, 122, 129, 138–40, 146, 150, 162, 163, 165, 166, 169, 172–4, 179, 181, 182, 185, 190, 194, 196, 199, 201, 214, 236, 244, 245, 247–52, 259–62, 267–77, 286, 296, 302, 304, 305, 315, 316, 328, 336, 338
 - speech community, 152, 159, 160, 178, 249
 - spelling, 10, 30, 31, 38, 44, 49, 76, 80, 215, 216, 228–30, 232, 234
 - spontaneous speech, 73, 248
 - standard
 - standard language, 258n11
 - structured query language (SQL), 111
 - students
 - A-level, 200n19, 201, 205, 281
 - GCSE, 200, 266n2
 - MA, 245–62
 - postgraduate, 180, 185, 214, 216, 294, 295
 - school, 53, 150
 - undergraduate, 89, 112, 165, 180, 185, 214, 223, 227, 293–295
 - Stylene (Environment for Stylometry and Readability Research for Dutch), 262
 - Survey of English Dialects (SED), 100, 103, 104
 - Survey of English Usage (SEU), 4, 7, 13, 14, 211, 214, 215
 - sustainability, 8, 11, 13, 27, 71, 91–3, 102, 111, 126–8, 177, 179, 180, 189, 206
 - synchronic, 57
 - Syntactic Atlas of the Dutch Dialects (SAND), 101, 111n9, 254, 255
- T**
- tablet, 37, 124, 150, 217, 219, 236, 237
 - tag, 47, 190–4, 198, 200–2, 254–6, 258, 314
 - tagged, 140, 184, 192, 194
 - tagging, 202, 255

- The Talk of the Toon, 13, 177–207,
268, 286
- target variety, 268
- teachers, 10, 15, 81, 84, 86, 89, 91,
125, 147, 150, 168, 194,
198, 200, 201, 203–5, 212,
221, 222, 234, 251, 266,
267, 269, 270, 274, 278,
279, 281, 285–8, 336, 337,
339, 341, 342
- training, 14, 15, 285
- textbooks, 90, 135, 148, 213–15,
222, 236, 269, 271, 336,
341
- Text Encoding Initiative (TEI), 36,
40, 46–51, 61, 62, 180,
181, 189–93, 198, 206, 207
- text messaging (SMS), 32
- time-aligned transcript, 137, 139
- time anchor tag, 190, 198, 200
- topic control, 305
- topic management, 303
- training materials, 342
- transcribed corpus, 268
- transcript/transcription
- conventions, 141, 304, 315–16,
335
 - protocol, 109
- translation task, 254
- turn, 71, 108, 111, 113, 213, 215,
221, 291, 293, 304–9, 314,
315, 322, 341, 342
- turn-taking
- latched turn, 307, 315
 - latching, 140, 293
- Twitter, 80, 81, 83, 87, 88, 93, 94,
123
- TXM, 189n14
- Tyneside Linguistic Survey (TLS), 6,
181–5, 192, 193, 204
- U**
- UK Data Service, 278
- United States of America
- Atlanta, 160, 169
 - North Carolina, 137
 - Ocracoke, 137
 - Roswell, GA, 12, 160, 162
- user comments, 122
- utterance, 16, 139–41, 190, 194,
196, 199, 285, 291, 292,
296, 305, 310, 311,
314–16, 337
- V**
- variation, 12, 15, 29, 37, 38, 100,
101, 103, 113, 129, 133,
134, 143, 144, 146, 152,
181–3, 187, 200, 226,
244–6, 249–50, 252–5,
259, 260, 262, 265, 266n2,
267, 285, 286
- video, 37n13, 73, 90, 124, 135,
144–8, 150, 187n12, 195,
203, 204, 207, 267, 295,
312, 336, 341
- vocabulary, 38, 82, 99, 100, 103–5,
107, 113, 118–21, 123,
182, 204, 212, 225, 228,
235
- Voices of Migration and Return
(VMR), 27

W

Wales, 200n19, 265
WAV, 109, 117, 127
The Web as Corpus (WebCorp), 233
web form, 93
Wmatrix, 297, 298, 301
word list, 186
Wordsmith/Wordsmith Tools, 36,
37n12

workshops, 3, 42, 48n20, 93,
200, 204, 205, 269,
285–7, 322, 326–9, 331,
332
written reports, 322, 327, 332

X

Xaira, 189n14