1

Taming Digital Voices and Texts: Models and Methods for Handling Unconventional Diachronic Corpora

Joan Beal, Karen Corrigan and Hermann Moisl

1 Stimulus for the volume and its overarching aim

Four of the contributions to Volume 2 (Allen, Beal, Corrigan, Maguire and Moisl; Standen and Jones; Meurman-Solin; and Raumolin-Brunberg and Nevalainen) arose from invited presentations at the workshop on 'Models and Methods in the Handling of Unconventional Digital Corpora' organized by the editors of the present volume that was held in April 2004 during the Fifteenth Sociolinguistics Symposium (SS15) at the University of Newcastle. The book project then evolved by inviting further contributions from key corpus creators so that the companion volumes would contain treatments outlining the models and methods underpinning a variety of digitized diachronic and synchronic corpora with a view to highlighting synergies and points of contrast between them. The overarching aim of the project is to establish whether or not annotation standards and guidelines of the kind already employed in the creation of more conventional corpora on standard spoken and written Englishes, such as the British National Corpus (http://info.ox.ac.uk/bnc) and the Bank of English (http://titania.cobuild.collins.co.uk/boe_info.html), should be extended to less conventional corpora so that they too might be 'tamed' in similar ways.

Since the development of the Brown corpus in the 1960s (see Francis and Kučera, 1964), the variety of electronic corpora now available to the linguistics community and the analytical tools developed to successfully mine these data have gone hand in hand with improvements

in standards and guidelines for corpus creation and encoding. Historical and vernacular electronic corpora of the kinds described in Volume 2 pose an array of additional problems as regards standards, since the creation of such databases often requires the encoder to come to the task *ab initio*. As such, while the resultant corpora are clearly high quality resources in their own right (and extremely valuable research tools within the discipline to which they relate), there is considerable variation in the models and methods used in the collection of these digital corpora and in their subsequent encoding and analysis, largely because the underlying theoretical goals and assumptions of the researchers are quite distinctive (cf. Ochs, 1999; McEnery and Wilson, 2001; Milroy and Gordon, 2003, p. 143; section 2.2). There are marked differences, for instance, in the nature of the data contained therein and they also vary in: (i) the levels of phonetic, lexical, grammatical and semantic annotation that they encode; (ii) the manner in which information is accessed/retrieved by the end-user and the manner in which it is displayed (whether or not the written/spoken word or multilingual texts are aligned, for example).

Advances in technology, from the ability to digitize historical manuscript materials and field recordings to the dramatic improvements in computer hardware, software, storage facilities and analytical tools, have enabled the collection and organization of such data sets into a growing number of user-friendly electronic corpora. The latter have the potential to offer new insights into linguistic universals, for instance, since they allow, for the first time, rapid, systematic and efficient comparisons to be made between languages across both time (real/apparent) and space (geographical). In addition, these corpora should be utilizable by researchers from a range of disciplines so that they are potentially as accessible to the socio-phonetician as they are to the discourse analyst or historical linguist in keeping with the aspirations of the Linguistic Data Consortium and Oxford Text Archive, inter alia.

These companion volumes are unique, since public output to date has primarily concentrated on describing and assessing the models and methods which underpin conventional corpora and the annotation standards/analytical tools developed specifically for them.[1]

2 Outline of contributions and their methodologies

Will Allen, Joan Beal, Karen Corrigan, Warren Maguire and Hermann Moisl in their chapter discuss the issues which arose in the compilation of the Newcastle Electronic Corpus of Tyneside English (NECTE). This

corpus differs from all the others described in this volume in various ways, not least of which is that the primary data from which it is compiled consist of spoken rather than written material. As a corpus of twentieth-century English, consisting of data recorded in 1969 and 1994, but including speakers born as early as 1889, it covers a period which has only very recently come to the attention of historical linguists (Mair, forthcoming). Like the corpora described in the chapters by Tagliamonte and by Anderwald and Wagner (both Volume 1), the NECTE corpus consists of data from a non-standard variety of (English) English. However, the NECTE corpus is more complex in that it makes available to researchers data in a number of formats: orthographic transcription, phonetic transcription and sound files, as well as providing TEI headers which include detailed social information about the speakers and the circumstances in which they were recorded. The team compiling the NECTE corpus set out both to preserve and make available data from two sociolinguistic surveys of Tyneside, both groundbreaking in their own ways: the Tyneside Linguistic Survey (Strang, 1968) and *Phonological Variation and Change in Contemporary Spoken English* (Milroy *et al.*, 1997). The methodologies of these two surveys, and the difficulties encountered in attempting to trace and preserve all materials from the former, are described here.

Other chapters in this volume discuss the challenges posed by attempts to provide 'diplomatic' editions of early texts, in which spelling and punctuation are highly variable. The transcription of spoken, regional data throws up similar problems, since dialect words often have no fixed spelling: indeed, some have never been encountered in print before. This chapter outlines the procedures which were followed by the NECTE team in reaching decisions about transcription and in creating a glossary.

Another challenge faced by the NECTE team was that of tagging the orthographic transcription files: at the outset they were pessimistic about the suitability of automatic tagging programmes designed for use with Standard English texts. However, it did prove possible to adapt CLAWS4 for this purpose.

The NECTE corpus is a Text Encoding Initiative (TEI)-conformant XML document, the standard recommended by the Arts and Humanities Data Service. As van Bergen & Denison (this volume) also point out, potential end-users are not all familiar with these standards, so the next phase of NECTE will involve the creation of style sheets which will convert the NECTE corpus into HTML and plain text versions for ease of visualization.

Susan Fitzmaurice's chapter describes a diachronic corpus designed not to be representative of the language of the period covered (1653–1762), but rather to provide a sample of the written repertoire of a network of individuals living between 1631 and 1762, with letters from a number of contemporary writers unconnected with this network included as a control. As such, it demonstrates, alongside Raumolin-Brunberg and Nevalainen (this volume) the ways in which diachronic corpora can be used for investigations in the emerging discipline of historical sociolinguistics (Nevalainen and Raumolin-Brunberg, 2003). In particular, the Network of Eighteenth-Century English Texts (NEET) is designed to address issues raised by the application of social network theory (Milroy, 1987) to historical (socio) linguistics (Tieken-Boon van Ostade, 1996, 2000a, 2000b), and a number of sample analyses are provided in this chapter. Texts included in this corpus come from four genres: letters, essays, fiction and drama, but it is the letters which pose the greatest challenges with regard to methods of corpus creation and annotation. Although searchable text cannot preserve the visual character of autograph manuscripts, the corpus preserves as far as possible the spellings, abbreviations, deletions and insertions found in the originals. However, this faithfulness to the original is at the expense of accessibility, so, in order to enable tagging, parallel, modernized text files had to be produced for some of the texts in the corpus. This need for parallel versions of texts when these are to be subjected to different levels of analysis is also a feature of the NECTE corpus and the LIDES project (Allen *et al.*, this volume, and Gardner-Chloros *et al.*, Volume 1). Even with modernized text as input, Fitzmaurice found that the automatic tagger designed for use with modern American English had to be substantially adapted in order to facilitate analysis of eighteenth-century texts. Once again, there is a parallel here with the way in which the CLAWS tagger had to be adapted in order to work with a regional variety of twentieth-century English (Allen *et al.*, this volume). Like the other corpora discussed in this volume, the NEET corpus holds potential interest for scholars from a range of disciplines, including historical and literary studies. Fitzmaurice points out here the importance of including as much historical, textual and bibliographical information as possible in the header, in order to facilitate access. Like Raumolin-Brunberg and Nevalainen (this Volume), she notes that, when creating headers for diachronic corpora, matters such as social classification are not straightforward, since any such categorization depends on knowledge of the ranks and strata of society in the eighteenth century.

Elizabeth Gordon, Margaret Maclagan and Jen Hay describe the Origins of New Zealand English (ONZE) corpus, which contains recordings of people born in New Zealand from the 1850s to the 1980s, that is, from shortly after the official start of the European settlement of New Zealand in 1840. It consists of three separate collections: (i) the Mobile Unit (MU) archive of speakers born between 1851 and 1910, (ii) the Intermediate Archive (IA) of speakers born between 1890 and 1930, and (iii) the Canterbury Corpus (CC) of speakers born between 1930 and 1984, which began in 1994 and has been added to each year since. The ONZE project was originally set up to study the process of new dialect development in New Zealand using the MU corpus, but is now also using the other two as well for both diachronic and synchronic study of New Zealand English.

The chapter is in four main parts. As its title indicates, the first part, 'Format and purpose of the archives', gives a detailed account of the above component parts of the ONZE corpus, and raises methodological issues specific to each: for MU reorganization of spoken material and identification of speakers, for IA dealing with different types of recordings, and for CC confidentiality with respect to still-living speakers. The second part, 'Preparation of the data', describes the reworking of MU, IA and CC by the ONZE project via re-recording onto modern media and orthographic transcription of the recordings. It also gives details of how information about speakers is held in databases, and refers to work which is under way to integrate this information with relevant analytical results. The third part, 'Making the data available to researchers', first deals with copyright issues, and then outlines several measures 'towards a digital interactive format for the corpus' (p. 00): time-alignment of audio and associated transcriptions using the Transcriber application, conversion to Praat-compatible format to enable acoustic analysis, and development of a web-based application for viewing, filtering and searching the time-aligned corpus. Further developments are projected for streamlining the current technicalities of accessing the desired part of the corpus and to go beyond orthographic to phonological search functionality. The fourth and final part, 'Use of the corpus', outlines the types of analysis that have been carried out on the corpus to date. Since the original aim of the ONZE project was to trace the development of the New Zealand accent, the initial analyses were phonetic/phonological and, more specifically, were of three types, each of which is briefly described: auditory perceptual, auditory quantitative and acoustic. The authors note that the digital formatting of the material will enable 'a much broader

perspective on sound change than has so far been possible' (p. 00): grammatical analyses beyond the phonological level, as well as extraction of information of interest to non-linguistic researchers like historians. A Conclusion summarizes the discussion.

Raymond Hickey's chapter describes a diachronic textual corpus of Irish English spanning 600 years from the early fourteenth century to the twentieth. As such, it provides a useful historical overview for the variety tackled from a purely synchronic perspective by Kallen and Kirk in their contribution to Volume 1. The Hickey corpus also relates to other atlas-type projects such as the Syntactic Atlas of Dutch discussed in Volume 1 by Barbiers *et al*.

A Corpus of Irish English (originally published as Hickey, 2003) aimed to collect a sample of Irish English texts that could be analysed so as to further our understanding of the genetic development of Irish English (Filppula, 1999) and to assess the impact that this variety has had on extraterritorial Englishes (Tagliamonte, 2000–03; Hickey, 2004).

As with many corpus creators in these volumes, Hickey addresses issues of representativeness, which, in his particular case, revolved around selecting texts from a range of periods and concentrating on those that were 'linguistically representative of Irish English' irrespective of whether or not they had 'literary merit' (p. 00). There was also an attempt to favour dramatic texts over other kinds, since there is an argument that these may well approximate speech patterns more readily and thus be less constrained by what Milroy (2001, p. 535) has termed 'the standard language culture' more often associated with written materials.

This chapter additionally offers preliminary analyses of the corpus, addressing the critical question in historical (socio-) linguistic research of how to make the best use of problematic data containing lacunae of various sorts (see Labov, 1994, p. 11) and the extent to which these limitations can be overcome by taking a 'careful and objective' (p. 00) approach to analysing such data.

Anneli Meurman-Solin describes the Corpus of Scottish Correspondence (CSC), whose creation was primarily motivated by the realization that 'royal, official, and family letters were a data source with unique properties in research seeking the reconstruction of both past language use and social as well as cultural practices ... Correspondence is a unique source in the sense that it offers both linguists and historians a wide range of informants representing different degrees of linguistic and stylistic literacy and different social ranks and mobility' (p. 00). Since the Corpus of Early English Correspondence,

described in this volume, covers East Anglia, London and the North of England, 'the focus on Scotland seemed very relevant' (p. 00).

CSC is based on diplomatic transcripts of the original manuscripts and is continually being expanded as transcription, digitization and tagging proceed. Revised and expanded versions will be distributed annually; the first distribution will comprise about 500,000 words of running text. The description of CSC is in four main parts. The first, 'Representativeness', outlines the criteria for the selection of texts to include in the corpus: (i) to ensure diachronic and diatopic representativeness, 'so that the corpus will permit the creation of a diachronic linguistic atlas and provide data for historical dialectology' (p. 00), (ii) positioning of texts on a continuum in accordance with their validity as evidence for particular research questions, and (iii) inclusion of 'variables relevant in the framework of historical sociolinguistics and historical stylistics and pragmatics' (p. 00). These criteria are discussed in detail. The second part, 'Digitization principles', notes that the principles used in digitizing the diplomatic transcripts of the archival materials will be described in the forthcoming corpus manual, but provides a brief summary of general practices with regard to such things as change of hand, folio number, paragraph structure, and so on. The third part, 'Basic and elaborated tagging', gives an account both of the principles governing the tagging of the CSC and of their implementation. The tagging is, first, 'designed to reflect a profoundly variationist perspective. The shape of Scots over time, place and social milieu is assumed to reflect continued variation and variability, resulting in a high degree of language-internal heterogeneity' (p. 00); an essential of the tagging system is that it should enable identification and analysis of complex patterns of variation and tracing of multidirectional processes of change. Second, the tagging system 'has been tailored to meet the challenge of tracing developments over a long time span' (p. 00). Third, the system must accommodate 'the inherent fuzziness and polyfunctionality recorded in language use when examined drawing on representative large-scale corpora' (p. 00). Fourth, the system must allow for the full range of zero realizations of grammatical features included in variationist paradigms. And, finally, the tagging system must 'provide information about the non-linguistic features of the original manuscripts whenever such information may affect analysis and interpretation of linguistic features' (p. 00). The tagging software that implements these criteria and used by CSC is agnostic with respect to 'modern formal syntactic theory' (p. 00), and provides annotation both at the level of word and morpheme and at higher-level syntactic

and discourse units. There follows a very detailed account of the tagging scheme. The fourth and final part of the discussion, 'Work in progress', gives a brief indication of what its title suggests.

Helena Raumolin-Brunberg and Terttu Nevalainen's chapter describes another corpus, or, rather, suite of corpora, designed for a very specific purpose, in this case, as with the NEET corpus, to test the methodology of historical sociolinguistics. The time span of the material included in these corpora is from 1410 to 1681, thus overlapping with the earlier part of the NEET corpus. The Corpus of Early English Correspondence (CEEC) corpora consist of the original 1998 corpus of 2.7 million words; the CEEC Sampler, a more accessible subcorpus of the 450,000 words not subject to copyright restrictions; the CEEC Supplement, consisting of material either not available in 1998, or only available in modernized spelling; the CEEC Extension, consisting of later material from 1681 to 1800; and the Parsed Corpus of Early English Correspondence. In this chapter, the authors concentrate on discussing issues which arose in the compilation of the original CEEC and the CEEC Sampler.

Although, like the Corpus of Late Eighteenth-Century Prose and parts of the NEET corpus, the CEEC corpora consist of letters, issues of transcription are not so important because CEEC consists of letters from edited collections rather than manuscript versions. The compilers are therefore dependent on the original editors, some of whom were historians rather than linguists, and so cannot always be sure that the 'diplomatic' text aimed at by Fitzmaurice and by van Bergen and Denison (this volume) has been achieved. However, a coding is provided to alert the user as to the extent of 'authenticity' of the text, from A ('autograph letter in good original-spelling edition') to D ('doubtful or uncertain authorship; problems with the edition, the writer's background, or both') (p. 00).

The advantage of using edited collections is that, for the most part, the texts could be scanned in, allowing for a much larger corpus to be compiled in the time available. Given the team's intention of testing sociolinguistic methods on historical data, this is important, as a large number of informants would be needed if cells containing speakers sharing social attributes such as gender, age and social level were to be filled for successive historical periods. The chapter contains a discussion of the difficulty of providing a 'balanced' sample from letters, when literacy was much less common amongst women and the lower social orders.

One disadvantage of using edited rather than manuscript materials is that issues of copyright arise. As the A5thors explain, this is not a

problem when a corpus is intended only for private research, but gaining copyright clearance becomes a major task if the corpus is to be widely accessible. The creation of a smaller corpus of texts out of copyright (the CEEC Sampler) has provided an interim solution to this problem.

There is considerable discussion of the problems encountered in coding and the solutions arrived at. As was the case for the NEET and NECTE corpora, the compilers of the CEEC found that automatic parsing programs designed for use with (Standard) present-day English were not suitable for use with early texts. In this case, the Penn Treebank program (see Taylor, this volume) proved successful. The authors also describe their solution to the problems posed by the need to include a wide range of background information on the letter-writers, given that the corpus was designed for use in socio-historical investigations. The authors argue that neither the Cocoa format used in the Helsinki Corpus of English Texts, nor the TEI model used in the NECTE corpus (Allen *et al.*, this volume) would allow the user to conduct searches of the data within combinations of parameter values. They therefore decided to create a database of social information on the senders of letters which could be searched separately.

The authors conclude with an overview of research which has made use of the CEEC corpora. What is evident here is that, by conducting pilot studies from an early stage of the project, as reported in Nevalainen and Raumolin-Brunberg (1996), the team has been able to use the results from these studies to inform the principles of compilation. Like most of the corpora discussed in this volume, the CEEC is a work in progress rather than a 'once and for all' finished article.

Naomi Standen and Francis Jones describe a project 'which will create and store in a database English translations for a set of five interlinked histories written in China between 974 and 1444 CE' and, when completed, 'will form an invaluable resource for historians' (p. 00). The authors are aware that, in describing a corpus intended primarily for historical study, their work appears to sit uncomfortably in a volume devoted to the creation of corpora for linguistic and more specifically sociolinguistic research. They point out, however, that translation from one language to another 'like all linguistic communication, has sociolinguistic significance': 'we use translation as an analytical tool to highlight the evolving relationships between terms and concepts in the Chinese originals' (p. 00). Their chapter develops this sociolinguistic dimension of the project in aiming to 'track the linguistic socio-ethnography of our own interpretative processes as translator-historians

within a framework of cross-border and post-colonial power relations' (p. 00).

The discussion is in four main parts. The first part, 'The database project: issues and aims', describes the historical context, role and nature of the Chinese texts to be translated, paying special attention to the complex textual and conceptual interrelationships of their narratives, and to the challenge of designing the translation database in a way that 'retains an openness to the multiple readings of events generated by the various source texts and their translations' (p. 00). The title of the second part, 'Ideology, history, and translation', gives a good indication of its content. Starting from the assumption that 'communication involves interpretation', and extending it to the observations that 'historians do not describe "what happened", but give their own reading of data' and that, because no two languages have identical grammatical structures, 'translators, too, give their own reading of the textual and paratextual evidence of their source text' (p. 00), the authors discuss such issues as 'terminology and attitudes' (the concept of barbarianism and its ideological application), 'translation as closure' ('how translation can fix and conceal the ideological subtexts inherent in any historical reading', p. 00), 'translation as opening' (ways of avoiding translational closure in the preceding sense), and 'terminology control and multiple meanings' (terminological standardization and its relation to translational openness and closure). The third part of the discussion, 'Technical solutions', outlines the design and high-level implementation of the project as (i) a collection of interlinked passages, (ii) a relational database of the links to permit the standard relational search operations on the corpus, and (iii) a glossary 'which will chiefly document multiple English translations of a single Chinese word' (p. 00). Finally, the fourth part, 'Creating the database', describes procedural aspects of the creation of the corpus. The Conclusion points out the fruitfulness of their 'methodological synergy between the "core" discipline of historical textual analysis on the one hand, and translation-studies approaches to textual transformation on the other' (p. 00), and suggests that similar approaches in other cases of complex intertextuality or difficult-to-align texts might prove useful in other applications, citing in particular the York–Helsinki corpus described in this volume.

Ann Taylor describes the York–Toronto–Helsinki Parsed Corpus of Old English Prose (YCOE), a 1.5 million-word syntactically annotated corpus of Old English prose texts produced at the University of York in the UK from 2000 to 2003 by Ann Taylor, Anthony Warner, Susan

Pintzuk and Frank Beths. The YCOE is part of the English Parsed Corpora Series. It is the third historical corpus to be completed in this format and follows the same kind of annotation scheme as its sister corpora, the Penn–Helsinki Parsed Corpus of Middle English II and York–Helsinki Parsed Corpus of Old English Poetry. In addition, two other corpora of the series, the Penn–Helsinki Parsed Corpus of Early Modern English and the parsed version of the Corpus of Early English Correspondence, are currently under construction at the University of Pennsylvania in the USA and the University of York in cooperation with the University of Helsinki, respectively.

The description is in five parts. The first part, 'Background', outlines the motivation for the creation of YCOE. The discussion begins with the observation that the corpus series to which YCOE belongs 'was designed particularly with historical syntacticians in mind, and more particularly, those who use quantitative methods in their work' (p. 00), and goes on to develop the specific need for syntactically annotated electronic corpora. The argument, in brief, is that, relative to paper-based corpora, electronic corpora offer well-known advantages of accessibility and amenability to fast and reliable computational analysis, but that 'virtually all the questions that interest syntacticians require structural information about language that is not accessible from word strings' (p. 00), and that this necessitates the insertion of grammatical tags, thereby making the general advantages of electronic corpora available to syntactic analysis. The subsection 'Research applications' identifies the main research uses for YCOE: 'studies of the sentential syntax of the various stages of English, either synchronic or diachronic' and more generally 'any sort of syntactic study, as well as many morphological studies' (p. 00). The second part, 'Methodology and representation', describes the content of YCOE (a subset of the 3,037 texts in the Old English corpus created for the *Dictionary of Old English*, using complete texts rather than samples), how these texts were formatted for part-of-speech tagging by the Brill system, the error correction procedure, and, finally, automatic parsing into Penn Treebank format. The third part, 'Structure', gives a detailed account of the YCOE annotation scheme and the principles that underlie it. The fourth part, 'Distribution and end-user issues', deals with availability of the text of YCOE and of documentation for it, and describes the features of CorpusSearch, an analytical tool developed in part by the author that will search 'any corpus in the correct format, including all corpora in the English Parsed Corpora Series' (p. 00). The Conclusion briefly summarizes the significance of YCOE in 'the programme of

creating syntactically parsed corpora for the whole attested history of the English language' (p. 00).

Linda van Bergen and David Denison's chapter describes the genesis of a relatively small (300,000-word) corpus of unedited letters, designed from the outset to be of interest to non-linguists, particularly historians, as well as linguists. Like the NEET corpus (Fitzmaurice, this volume), the Corpus of Late Eighteenth-Century Prose (CLEP) plugs the gap between the major corpora of earlier English and larger modern corpora such as the British National Corpus. The authors here describe their corpus as 'opportunistic in origin' (p. 00), presumably because the material was available in the archive of the John Rylands University Library in Manchester, but, as such, it provides an example of the kind of project which could be replicated with material from archives elsewhere in the UK. Van Bergen and Denison discuss in detail the decision-making processes involved in selecting material for their corpus, which consists of letters written to Richard Orford, a steward to the Legh family of Lyme Hall, Cheshire. Unlike the letters included in NEET, which were written by and to literary figures, and chosen to facilitate investigations into standardization and the effects of prescriptivism, those in CLEP represent the 'everyday' language of informal business transactions. Although the purpose of the letter-writers is primarily to conduct business, personal matters often intrude, so that, as the authors state, 'the dividing line between business and personal letters turned out to be very fuzzy' (p. 00). The decision to include rather than exclude material has led to what the authors admit is an 'unbalanced and heterogeneous' (p. 00) corpus, but, as with other corpora making use of historical and/or archive data (see, for instance, Allen *et al.*, this volume), the richness of the data included in the corpus leads us to question whether 'balance' is after all essential in diachronic corpora.

With regard to transcription, van Bergen and Denison report that, like Fitzmaurice (this volume), they aimed for a 'diplomatic' edition of the text, that is, one as close as possible to the original. The inclusion in this chapter of illustrations of the actual manuscript for comparison with the 'diplomatic' texts allows us to judge the closeness of the latter to the originals. However, as was the case for the NECTE corpus (Allen *et al.*, this volume), transcription was not straightforward, in this case because the handwriting was not always easy to decipher, so tentative or dubious readings are marked as such.

The discussion of coding in this chapter foregrounds an issue which was extensively debated in the workshop at the Fifteenth Sociolinguistics

Symposium with which these volumes are closely associated. The compilers of CLEP decided to use coding based on that of the Helsinki Corpus, because most users would be familiar with this. The corpus is available in two versions: a plain text file for concordancing and an HTML version designed for use with web browsers. There is an illustrated account of the advantages of HTML, followed by a discussion of the pros and cons of using TEI-conformant coding such as XML (as used in the NECTE corpus). In this case, the authors argue that potential users are not sufficiently familiar with XML to make this a practical option for the first release.

3 Acknowledgements

The editors would like to close by acknowledging the financial support provided to various phases of this project by: (i) the School of English Literature, Language and Linguistics, University of Newcastle; (ii) Newcastle Institute for the Arts, Humanities and Social Sciences; (iii) Palgrave Macmillan; and (iv) the Arts and Humanities Research Council (grant no. RE11776).

We would also like to express our deeply felt gratitude to our authors who have gracefully endured our cajoling and actively engaged with us in pursuing a research agenda in corpus linguistics, the ultimate goal of which is to foster 'international standards for metadata', and articulate 'best practices for the collection, preservation, and annotation of corpus data for language archives' (Kretzschmar *et al.*, 2005 and forthcoming).

There are also a number of other people who deserve special thanks, including: Jill Lake, the commissioning editor responsible for these companion volumes, for her helpful feedback from inception to completion; Melanie Blair, editor of the series, for her patience with our many technical queries; Tina Fry, Alison Furness, Kaycey Ihemere and Adam Mearns for their assistance with formatting and indexing; the organizing and scientific committees of SS15 and our anonymous reviewers who submitted the SS15 workshop papers and the chapters in these volumes to critical, stylistic and formal scrutiny. Finally, we are indebted to Shana Poplack for writing the foreword to this volume and for the many discussions we have had with her regarding the shape that these volumes should take since the idea for this project was first mooted back in 2002. Any remaining shortcomings are, as usual, our own.

Note

1. See, for instance, Francis and Kučera (1964); Johansson *et al*. (1978); Aarts and Meijs (1984); Garside (1987); Garside *et al*. (1987); Leech (1992); Hughes and Lee (1994); Burnard (1995); Haslerud and Stenstrom (1995); Sampson (1995); Knowles *et al*. (1996); Aston and Burnard (1998); Biber *et al*. (1998); Condron *et al*. (2000), inter alia.

References

Aarts, Jan and Willem Meijs (eds). 1984. *Corpus Linguistics*. Amsterdam: Rodopi.

Aston, Guy and Lou Burnard. 1998. *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Burnard, Lou. 1995. *Users' Reference Guide to the British National Corpus*. Oxford: Oxford University Computing Services.

Condron, Frances, Michael Fraser and Stuart Sutherland. 2000. *Guide to Digital Resources in the Humanities*. Oxford: Humanities Computing Unit, Oxford University.

Filppula, Markku. 1999. *The Grammar of Irish English*. London: Routledge.

Francis, W. Nelson and Henry Kučera. 1964. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Providence, RI: Dept. of Linguistics, Brown University.

Garside, Roger. 1987. 'The CLAWS word-tagging system'. *The Computational Analysis of English: A Corpus-Based Approach*, ed. by Roger Garside, Geoffrey Leech and Geoffrey Sampson, pp. 30–41. London: Longman.

Garside, Roger, Geoffrey Leech and Geoffrey Sampson (eds). 1987. *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman.

Haslerud, Vibecke and Anna-Britta Stenstrom. 1995. 'The Bergen London Teenage Corpus (COLT)'. *Spoken English on Computer*, ed. by Geoffrey Leech, Greg Myers and Jenny Thomas, pp. 235–42. London: Longman.

Hickey, Raymond. 2003. *Corpus Presenter. Processing Software for Language Analysis with a Manual and A Corpus of Irish English as Sample Data*. Amsterdam: John Benjamins.

Hickey, Raymond (ed.). 2004. *Legacies of Colonial English: Studies in Transported Dialects*. Cambridge: Cambridge University Press.

Hughes, Lorna and Stuart Lee (eds). 1994. *CTI Centre for Textual Studies Resources Guide 1994*. Oxford: CTI Centre for Textual Studies.

Johansson, Stig, Geoffrey N. Leech and Helen Goodluck. 1978. *Manual of Information to Accompany the Lancaster–Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Dept. of English, University of Oslo.

Knowles, Gerry, Briony Williams and Lolita Taylor. 1996. *A Corpus of Formal British English Speech*. London: Longman.

Kretzschmar, William A., Jean Anderson, Joan C. Beal, Karen P. Corrigan, Lisa-Lena Opas-Hänninen and Bartek Plichta. 2005. 'Collaboration on corpora for regional and social analysis'. Paper presented at AACL 6/ICAME 26, University of Michigan, Ann Arbor, 12–15 May 2005.

Kretzschmar, William A., Jean Anderson, Joan C. Beal, Karen P. Corrigan, Lisa-Lena Opas-Hänninen and Bartek Plichta. (forthcoming). 'Collaboration on corpora for regional and social analysis'. Special Issue of *Journal of English Linguistics*.

Labov, William. 1994. *Principles of Linguistic Change: Volume 1, Internal Factors*. Oxford: Blackwell.

Leech, Geoffrey N. 1992. '100 million words of English: the British National Corpus'. *Language Research* 28(1):1–13.

Mair, Christian. (forthcoming). *Standard English in the Twentieth Century: History and Variation*. Cambridge: Cambridge University Press.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics*, 2nd edn. Edinburgh: Edinburgh University Press.

Milroy, James. 2001. 'Language ideologies and the consequences of standardization'. *Journal of Sociolinguistics* 5(4):530–55.

Milroy, Lesley. 1987. *Language and Social Networks*. Oxford: Blackwell.

Milroy, Lesley and Matthew Gordon. 2003. *Sociolinguistics: Method and Interpretation*. Oxford: Blackwell.

Milroy, James, Lesley Milroy and Gerard Docherty. 1997. 'Phonological variation and change in contemporary spoken British English'. ESRC, unpublished Final Report, Dept. of Speech, Newcastle University.

Nevalainen, Terttu and Helena Raumolin-Brunberg. 1996. *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam and Atlanta, Ga: Rodopi.

Nevalainen, Terttu and Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Longman Linguistics Library. London: Longman.

Ochs, Elinor. 1999. 'Transcription as theory'. *The Discourse Reader*, ed. by Adam Jaworski and Nikolas Coupland, pp. 167–82. London: Routledge.

Sampson, Geoffrey. 1995. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon.

Strang, Barbara M. H. 1968. 'The Tyneside Linguistic Survey'. *Zeitschrift für Mundartforschung*, NF 4 (Verhandlungen des Zweiten Internationalen Dialecktologenkongresses), pp. 788–94.Wiesbaden: Franz Steiner Verlag.

Tagliamonte, Sali A. 2000–03. 'Back to the roots: the legacy of British dialects'. Unpublished report to the ESRC, grant no: R000239097.

Tieken-Boon van Ostade, Ingrid. 1996. 'Social network theory and eighteenth-century English: the case of Boswell'. *English Historical Linguistics 1994*, ed. by David Britton, pp. 327–37. Amsterdam and Philadelphia: Benjamins.

Tieken-Boon van Ostade, Ingrid. 2000a. 'Social network analysis and the history of English'. *European Journal of English Studies* 4(3):211–16.

Tieken-Boon van Ostade, Ingrid. 2000b. 'Social network analysis and the language of Sarah Fielding'. *European Journal of English Studies* 4(3):291–301.

Websites

Bank of English: http://titania.cobuild.collins.co.uk/boe_info.html
British National Corpus: http://info.ox.ac.uk/bnc