1

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

Joan Beal, Karen Corrigan and Hermann Moisl

1 Stimulus for the volume and its overarching aim

Six of the contributions to Volume 1 (Anderson *et al*.; Anderwald and Wagner; Barbiers *et al*.; Sebba and Dray; Kallen and Kirk; Tagliamonte) 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 this data have gone hand in hand with improvements in standards and guidelines for corpus creation and encoding. Contemporary spoken and written regional English corpora, as well as those containing bilingual and child language data of the kinds described in this volume, 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, section 2.2; Milroy and Gordon, 2003, p. 143). 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 first and second languages/dialects across genres as well as social and geographical space. In addition, these corpora should be utilizable by researchers from a range of disciplines so that they are potentially as accessible to the socio-syntactician as they are to the conversation analyst or child language specialist 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

The chapter by Anderson, Beavan and Kay provides an account of a corpus which is made up of data from a wide range of sources and in a number of formats: written text, audio and video. The SCOTS corpus is intended eventually to include data from all the languages spoken in Scotland, but at present consists of material in Scots and Scottish

English. The authors here acknowledge that these two varieties are best thought of as a continuum with 'broad Scots' at one end and (Standard) Scottish English at the other. As such, the SCOTS corpus fills the gap left by the lack of a Scottish component for the ICE corpus to match the ICE-Ireland corpus described by Kallen and Kirk in this volume.

The upsurge of interest in Scottish language and culture surrounding the devolution of political power from London to Edinburgh meant that the compilers of the SCOTS corpus had a very positive response to their initial call for material. At the time of writing, the corpus consisted of some 600,000 words, but, as a monitor corpus, it is updated whenever significant amounts of new data are available. The involvement of the public, and the open-ended nature of this corpus, meant that legal/ethical issues of copyright, along with practical considerations of record-keeping, are probably more critical than those which apply to many of the corpora in these volumes (though see Sebba and Dray, as well as Kallen and Kirk). The administration system developed in response to these challenges is described and illustrated here, providing a useful model for teams embarking on similarly complex projects.

Like the compilers of many other corpora described in these volumes, (for example, Allen *et al.*, Volume 2; Sebba and Dray, this volume) the SCOTS corpus team were confronted with the challenge of developing standards for transcribing texts for which standardized spelling conventions have not been developed. Indeed, they found that writers of some Scots texts submitted were not consistent in their spelling, even within the same text (the same issues were, of course, faced by the compilers of many of the diachronic corpora described in Volume 2). Since SCOTS is a searchable corpus, variability of spelling is also an issue for search-words: the team are, therefore, working with Scottish Language Dictionaries to devise headwords and a sophisticated enough spelling system.

Since the SCOTS corpus is intended to be publicly available to a wide variety of end-users, the website has been designed with accessibility in mind, using XHTML for the main website, but with provision for viewing and downloading plain text. The intention is to make TEI-compliant XML data sets available in future, in the manner of Allen *et al.*, Volume 2.

Finally, there is some discussion of the desirability of 'balance' in a corpus, but, as for most of the corpora described in these volumes, any attempt on the part of the SCOTS team to produce a corpus satisfying, for example, the British National Corpus (BNC) criteria for 'balance' and 'representativeness' would be futile, since, as the authors explain here, it is not yet known what proportion of writing in Scots is found in the various genres.

Lieselotte Anderwald and Susanne Wagner describe a corpus (the Freiburg English Dialect Corpus or FRED) which was compiled from data already collected by oral historians and, in that sense, there is an obvious relationship between it and NECTE (Allen *et al.*, Volume 2), though it does not share the latter's diachronic concerns. Consisting as it does of material recorded from older speakers in the 1960s and 1970s, it provides data which are comparable with those collected by the *Survey of English Dialects* (Orton, 1962), in that it represents the traditional dialect of speakers who reached adulthood before the Second World War, had minimal education and little or no mobility, and were mostly male. As such, FRED was designed to be regionally, rather than socially, representative.

The use of oral history data in a corpus digitized for linguistic purposes is novel, but, just as the diachronic corpora described in Volume 2 have proved of interest to historians, so linguists might well make more use of historical corpora.[2] Anderwald and Wagner discuss the drawbacks of using material compiled by non-linguists: transcription is often unsatisfactory, either 'normalizing' the text so that regional morphology disappears, or, conversely, using 'eye-dialect' in an attempt to convey the 'flavour' of the dialect. However, where audiotapes are also available, this can be corrected. A more serious drawback is the predominance of past over other tense forms, so that the corpus would not be useful for a study of present or future forms. Since the interviews are monologic, the FRED corpus would not be suitable for the study of discourse features, but the same could be said of much interview data.

Reports on preliminary investigations using the FRED data which close this chapter demonstrate that corpus-based studies of regional variation in the morphology and syntax of English are overdue: for instance, the view put forward by Wakelin (1975) that the dialect of West Cornwall is more 'Standard' than that of the rest of the South-West is contradicted by Wagner's study of gendered pronouns, which, in the FRED corpus, are, in fact, most common in this region.

The third chapter in this volume by Barbiers, Cornips and Kunst describes at length the 'taming' of a larger-scale dialect syntax project relying on more recent data, namely, the Syntactic Atlas of the Dutch Dialects (SAND). The survey was conducted in the Netherlands and

Flanders and it aimed to create both a traditional printed atlas and an electronic searchable version that used specially created cartographic software for generating maps of particular morphosyntactic features online.

The data upon which both atlases are based are drawn from a range of sources: oral interviews conducted at various sites and followed up by telephone as well as the more traditional postal survey method (though at 156,000 question–answer pairs, the SAND version of the latter is considerably more extensive than most[3]).

The chapter relates in considerable detail the methodologies that underpin this mammoth venture outlining the typical social characteristics of SAND informants as well as the specialized techniques used to elicit responses from them (and the advantages and disadvantages of these more generally).

Praat (a transcription tool originally developed for phonetic data) was used to create the electronic versions of the corpus as it allowed alignment between orthographic transcription and speech signal and also made it possible to signal different levels in the transcription (an often problematic issue with transcribed spoken corpora, which is naturally, therefore, addressed by other authors in Volume 1, such as Gardner-Chloros *et al*., as well as Allen *et al*. in Volume 2). The Barbiers *et al*. chapter also outlines the transcription protocol adopted, which, essentially, amounted to a normalization of the data into Standard Dutch orthography as far as practicably possible.[4] This was necessary for several reasons, most critical of which are: (i) it permits a certain amount of automatic lemmatization and pre-tagging and (ii) it has the added benefit that spelling across dialects is made uniform. The second of these is clearly an important consideration when you are developing online search and cartographic tools for a range of variables drawn from over 260 quite divergent dialects.

In contrast to most of the other linguistic corpora referred to in these volumes, SAND, like aspects of the SCOTS corpus described above, takes the form of a relational database. This permits more flexibility for the updating of the corpus in various ways over time and thus obviates the multiple versions issue that can arise when electronic corpora consist merely of collections of tagged text files. As already mentioned, the main function of SAND is to serve as a dynamic syntactic atlas and setting the corpus up as a relational database also facilitates its end-use as such in important ways. Users can perform database queries that can (both automatically and manually) create the input for maps that show the spatial distribution of syntactic variables across Flanders and The

Netherlands. Moreover, users can perform their own analyses by combining syntactic variables in one map so as to ascertain the extent to which their distribution coincides geographically. No doubt this unique corpus, which rests on meticulous methodological considerations, will live up to the aspirations of its creators, namely, to further research into syntactic microvariation from theoretical, typological, geolinguistic, historical and quantitative perspectives.

The next chapter, by Penelope Gardner-Chloros, Melissa Moyer and Mark Sebba, focuses on the Language Interaction Data Exchange System (LIDES) which is affiliated with the CHILDES enterprise, described below and in more detail in its creator's own chapter elsewhere in Volume 1 (see MacWhinney). The main goal of LIDES is to establish a network of scholars, working on the language interactions of adult multilingual speakers, who are committed to the same objectives with respect to developing coding schemes and guidelines for the electronic databases they produce.

Gardner-Chloros *et al*. in their chapter focus on the rationale behind the choice of the CHAT coding scheme and the CLAN tools in CHILDES for this purpose, noting that, despite some drawbacks, their adoption was encouraged by the fact that they are so open to further elaboration and updating and because: (i) there is a user-friendly interface already in place between CHAT and XML formats; (ii) CLAN programs recognize Unicode which is proving to be important for research on linguistic (including phonetic) data because it allows researchers to use their computer keyboard to represent a character from any language that has different script types;[5] and (iii) CLAN will eventually be fully able to support Praat which, as our discussion of the SAND project demonstrated, is increasingly becoming a standard tool for aligning and splicing real-time speech alongside written transcriptions of it.

In addition to facilitating the exchange and comparison of data sets created using these standards and tools, there are further benefits of LIDES, including the fact that the CLAN programs can be used to easily search very large data sets for patterns (of code-switching, for instance) or to provide quantitative analyses of various types.

A particular concern of the chapter is one shared by many contributors to these volumes, namely, the theoretical and practical problems encountered when transcribing spoken data. Indeed, Gardner-Chloros *et al*. clearly demonstrate that the difficulties described by our other authors whose corpora are composed of monolingual discourse are considerably exacerbated when transcribing *plurilingual* data. It is not surprising, therefore, that a considerable amount of discussion in the chapter revolves around the nature of the problems and the strategies which the LIDES team have found to be effective in resolving them. Critical to this dialogue is the working through of problematic cases using genuine data, a process which is concluded by presenting a step-by-step outline of the CHAT transcription scheme. The latter will clearly be invaluable to readers wishing to use the system to code their own multilingual data sets, as will the concluding section of the chapter in which applications (searches and frequency counts, for example) of CHAT-coded data are demonstrated.

Jeffrey Kallen and John Kirk, whose ICE-Ireland project has already been mentioned as sharing certain similarities with the SCOTS corpus in particular, provide the next chapter. It begins, naturally enough, with a brief history of the International Corpus of English (ICE) and its goals, and then turns to the specific application of these guidelines in an Irish context. At its most basic level, an ICE corpus must contain 300 spoken texts and 200 that are written, all of which are transcribed using the ICE coding system so that they can be compared with one another in a similar manner to that advocated in other international collaborative corpus projects featured in these volumes, such as CHILDES, LIDES and YCOE (see Taylor, Volume 2).

Although the criteria for creating the ICE corpora seem straightforward enough, a number of issues have had to be resolved by the ICE-Ireland team in their creation of an Irish version. An overriding issue (not faced by ICE-GB because of its southern-centric English bias but clearly also a consideration for other ICE collections, such as those in the Caribbean) was the definition of state boundaries, which Kallen and Kirk term the ‘national context issue’ (p. 00). Simply put, their concern is connected with the incongruity between more recent political and legal divisions of Ireland and its natural island geography as well as its more long-standing linguistic and social history, which do not necessarily match. The solution has been to eschew the production of two separate corpora (dictated by national boundaries) in favour of a single corpus that transcends these. This choice permits, for example, the inclusion of conversations between speakers both north and south of the technically separate state borders and a corpus content, which is split equally between Northern Ireland and the Republic. Although this distorts actual population distributions between the two jurisdictions, it is in keeping with practices adopted in the ICE programme more generally.

The collection and digitization of the ICE-Ireland corpus was begun in 1990 and has quickened in pace from 1999 onwards thanks to the receipt of research council and other awards. While additional material outwith the 1990–94 timescale of the ICE corpora more generally had to be collected in 2002 and 2003 to reach the target number of conversations and texts, the project is now complete and this chapter presents a wide-ranging account of the particular hurdles that had to be overcome in order to accomplish this. Of particular concern is the fact that access to certain kinds of data (such as recordings of courtroom proceedings and telephone conversations with male informants) normally found in ICE was impossible because of legal prohibitions and other culture-specific restrictions. Kallen and Kirk relate their attempts to fill these lacunae and the extent to which these were successful.

The constraints imposed by the transcription and annotation protocols of ICE meant that decisions with regard, for instance, to orthographic representation (a recurrent theme of chapters in each of these volumes) were considerably more straightforward for Kallen and Kirk, though they were also not without their own challenges (on account of having to encode certain unique aspects of Irish English in the digitized corpus).

Since so much of ICE-Ireland is prescribed in this way, the main focus of this chapter is on the kinds of analysis that can be performed on the finished product and therefore the sorts of research question it can be used to address. Given the nature of the corpus, these centre primarily on the extent to which Irish English is standardized (and thus similar to other regional standards digitized during the ICE programme) and the degree to which it retains regional dialect features (lexical and morphosyntactic) particular to the languages (Irish and English) of Ireland. Although their answers to these questions are naturally tentative at this stage in the research, ICE-Ireland offers intriguing data for the discussion of wider questions, such as the means whereby standard languages are also subject to structured patterns of variation and change more typically thought of as characteristic of non-standard dialects.

MacWhinney describes the current state and projected developments of the TalkBank project, ‘a new system that will lead to a qualitative improvement in social science research on communicative interactions’ (p. 00). He begins with the observation that phenomenal growth in the power and connectivity of computers and associated software developments have led to dramatic advances in the methodology of ‘hard’ science and engineering, but that the behavioural and social sciences have not shared fully in these advances due, in large part, to the complexity of human interactional behaviour and of the difficulty of representing this complexity in ways suitable for scientific analysis. TalkBank is a National Science Foundation-funded project whose goal 'is to support data-sharing and direct, community-wide access to naturalistic recordings and transcripts of human and animal communication' (p. 00), and which will address seven needs: (1) guidelines for ethical sharing of data, (2) metadata and infrastructure for identifying available data, (3) common, well-specified formats for text, audio and video, (4) tools for time-aligned transcription and annotation, (5) a common interchange format for annotations, (6) a network-based infrastructure to support efficient (real-time) collaboration, and (7) education of researchers to the existence of shared data, tools, standards and best practices.

The discussion looks at several issues in human behaviour and communications data analysis from a TalkBank point of view. The first issue is transcription of the 'complex pattern of linguistic, motoric, and autonomic behaviour' so as to 'capture the raw behaviour in terms of patterns of words and other codes' (p. 00). This has historically been difficult for three reasons: (1) lack of coding standards, (2) indeterminacy, that is, subjectivity and the possibility of error in representing what is observed, and (3) tedium, the labour-intensive nature of transcription and, again, the consequent probability of error. As a solution, TalkBank proposes digitization of the recording media and subsequent use of computational tools that provide time-aligned linkage of transcripts and codes with the original audio or video recordings. This relieves the transcriber of the need to represent as much of the original recordings as possible, since the analyst can check and where necessary augment the transcription.

The second issue is collaborative commentary. In text-based disciplines such as literary criticism and historical analysis, provision of an object of study and discussion of alternative views of that object have historically been the norm, and have now been made even easier via electronic text and communication media. Provision and discussion of spoken discourse have, until recently, been more cumbersome; MacWhinney and his co-workers have been developing an XML-based schema for making the CHILDES and TalkBank corpora available for collaborative commentary via the Web.

The third and final issue is 'community of disciplines': 'TalkBank seeks to provide a common framework for data-sharing and analysis of each of the many disciplines that studies conversational interactions' (p. 00). The data requirements of the following research areas are discussed: classroom discourse, animal communication, field linguistics, conversational analysis, gesture and sign, second language learning and bilingualism, aphasia, first language acquisition, cultural anthropology, psychiatry, conflict resolution, and human–computer interaction.

The discussion concludes with an outline of proposed further developments in TalkBank:

- Creation of a qualitative data analysis tool called Coder, which 'will allow the user to create and modify a coding framework which can then be applied to various segments of the transcript' (p. 00).
- Creation of new and flexible ways of displaying data.
- Development of more user-accessible ways of framing profiles and queries for data filtering.
- Provision of teaching tools that 'directly introduce students to the study of language behaviour and analysis' (p. 00).
- Transfer of control over construction of the TalkBank and associated data from the current few individuals to the relevant research community.

The next chapter, by Mark Sebba and Susan Dray, relates their experiences of developing two digital Creole corpora at Lancaster University, namely the Corpus of Written British Creole (CWBC) and its sister corpus the Corpus of Written Jamaican Creole (CWJC). Since each of these was created for a slightly different purpose, their annotation schemes are dissimilar in certain respects, though the basic principles behind them both are fundamentally the same. Their contribution addresses specific issues regarding the selection and annotation of English Creole texts, but it also draws out more general issues that are echoed in other chapters across both volumes that engage with 'unconventional' written texts. Their particular concern is the extent to which visual and graphological features contained in the originals need to be retained in the computerized versions.

The chapter opens by usefully defining and contextualizing the two kinds of Creole that were used in the creation of these corpora. As such, we are introduced to the very particular socio-historical circumstances that allowed British and Jamaican Creoles to develop and the exact relationship between them. A short history of writing in these varieties is also given, noting that the increased prestige of Jamaican Creole has impacted upon the range and number of written works in which it can be found. While British Creole does not benefit to the same extent from this increased acceptability, the number of public texts produced in this variety has grown from negligible in the 1980s to 'analysable' in the 1990s. As such, developing the CWBC provided an opportunity for researching the creation and exploitation of a very new form of unstandardized writing in English. The goal of the CWJC was to permit analyses of the distinctive non-standard writing strategies employed by Jamaicans so as to contribute to ethnographic research into this community's literary practices.

Given the contemporary nature of the written resources (published and unpublished) that comprise the CWJC and the CWBC, issues of copyright which beleaguered other corpus creators in this volume (SCOTS and ICE-Ireland especially) also applied. Since this project's financial resources were rather more meagre than that of the SCOTS corpus, for example, this restricted the size and composition of the CWJC and CWBC in various ways, as they were unable to develop a similar database system for pursuing permissions. By adapting a solution rather like that of Raumolin-Brunberg and Nevalainen (Volume 2) in relation to their Corpus of Early English Correspondence, Sebba and Dray were, nevertheless, able to produce, within a reasonable timescale, digital corpora containing various written genres of British and Jamaican Creole suitable for certain kinds of linguistic analysis.[6]

In order to facilitate analyses of various kinds, the corpora have been manually annotated using a set of contrastive tags marking differences in graphology and lexis as well as discoursal and grammatical structure between Standard English and the two types of Creole texts. Metadata of various types are also tagged with both geographical provenance and visual cues contained in the originals being preserved in this manner.

After a discussion of similar issues to those addressed by other authors in these volumes who have worked with 'texts' as opposed to 'voices' (the preservation of non-verbal meaning and the complex type–token relationship, in particular), the authors close by presenting some applications of the corpora. Although the relatively diminutive size of these corpora by comparison to SCOTS, SAND and ICE-Ireland, for instance, delimit the nature of linguistic investigations that can be performed on them, the CWBC has already been used for a small-scale study of future modality (Facchinetti, 1998) and for a study of orthographic practices (Sebba, 1998a, 1998b) which was able to discriminate genuine 'Creole' from 'non-Creole' texts. In the longer term, as a monitor corpus, the CWBC will become invaluable to socio-historical linguists for whom it will present a rare opportunity to document an unstandardized language developing written forms and functions.

The final chapter in Volume 1, by Sali A. Tagliamonte, focuses on what she considers to be 'tried and true procedures' (p. 00) for the creation and annotation of electronic corpora derived from vernacular speech. The chapter draws on Tagliamonte's own previous experience since the late 1980s of various corpus-building enterprises in order to illustrate the models which she advocates for the maximally efficient management and analysis of spoken data collected via the sociolinguistic interview method.

Tagliamonte's particular contribution in this regard is the very careful account that she gives of the importance of establishing strict protocols for orthographic transcription at the outset of any annotation programme. She argues that representing the non-standard phonologies and morphosyntactic features prevalent amongst speakers of distinctive social and regional varieties can be particularly problematic. It is therefore crucial that conventions regarding the kinds of information to encode are based on best practice and are strictly adhered to across the entire data set to allow maximum comparison between speakers across regional, social and temporal space.

In this regard, Tagliamonte is another advocate of the normalization of spelling and punctuation when digitizing corpora, though in this case her rationale is not primarily to facilitate database searches as it was for Barbiers and the SAND project team, but to enhance both readability and the speed of transcription (particularly key considerations when dealing with smaller-scale private corpus-building enterprises of this kind). Crucial too, of course, is the creation of orthographic protocols that do not normalize blindly, but incorporate specific unplanned aspects of speech as well as certain variant realizations that are so meaningful to your particular research questions that they are worth the extra effort to encode. Tagliamonte uses extracts from various corpora that she has been associated with to illustrate exactly what these might be and the manner in which they have been encoded for her particular purposes.

Again, as we have seen with all the corpora referred to in these two volumes, the data collection and annotation processes described in Tagliamonte's contribution go hand in hand with the creation of metadata of one sort or another (in her case a relational database created using the FileMaker program). This allows processing and searching of the data in various ways alongside the use of other software that can be applied to the text files themselves (such as Concorder), producing frequency counts per social category of speaker and so on. This then facilitates the transfer of the data to statistical packages that add further levels of encoding relating to the testing of research hypotheses of various kinds (GoldVarb is the one usually advocated in the kind of variationist research that Tagliamonte is associated with). The chapter concludes with a demonstration of these stages using genuine corpus data alongside a critical evaluation of the methods themselves.

Although the electronic corpora which Tagliamonte describes in this chapter are, like Dray's CWJC, 'private' in the sense of Bauer (2004), the discussion still has much to offer with regard to the principles of database management and the mechanics of corpus-building for all kinds of linguistic data sets.

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, Jill's assistant, 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.

Notes

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.
2. For a discussion of the advantages and disadvantages of using such resources as tools for linguistic analysis more generally, see Corrigan (1997).
3. Compare, for example, the relatively recent *Dictionary of American Regional English* (Cassidy and Hall, 1985–), http://polyglot.lss.wisc.edu/dare/dare.html, which was based on just over 1,800 questions (see Wolfram and Schilling-Estes, 2005, p. 126).
4. For instance, clitic clusters, for obvious reasons, were handled rather differently.
5. This is why Unicode was also preferred by the NECTE team (see Allen *et al*., Volume 2).
6. The type of analysis being restricted largely by the small sizes of the corpora (CWBC= 28,000 words and CWJC = 70,000 words), which presents issues of representativeness.

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.

Bauer, Laurie. 2004. 'Inferring variation and change from public corpora'. *Handbook of Language Variation and Change*, ed. by Jack K. Chambers, Peter Trudgill and Natalie Schilling-Estes, pp. 97–114. Oxford: Blackwell.

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.

Cassidy, Frederic G. and Joan H. Hall (eds). 1985–. *Dictionary of American Regional English*, 4 vols. Cambridge, Mass.: Harvard University Press.

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

Corrigan, K. P. 1997. 'The syntax of South Armagh English in its socio-historical perspective'. Unpublished doctoral thesis, National University of Ireland at University College Dublin.

Facchinetti, R. 1998. 'Expressions of futurity in British Caribbean Creole'. *ICAME Journal* 22:7–22.

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 Leech, Greg Myers and Jenny Thomas, pp. 235–42. London: Longman.

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*.

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

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

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

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

Orton, H. 1962. Survey of English Dialects: Introduction. Leeds: Arnold.

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

Sebba, M. 1998a. 'Meaningful choices in Creole orthography: "experts" and users'. *Making Meaningful Choices in English: On Dimensions, Perspectives, Methodology and Evidence*, ed. by R. Schulze, pp. 223–34. Tübingen: Gunter Narr.

Sebba, M. 1998b. 'Phonology meets ideology: the meaning of orthographic practices in British Creole'. *Language Problems and Language Planning* 22(1):19–47.

Wakelin, M. 1975. *Language and History in Cornwall*. Leicester: Leicester University Press.

Wolfram, Walt and Natalie Schilling-Estes. 2005. *American English: Dialects and Variation*, 2nd edn. Oxford: Blackwell.

Websites

Bank of English: http://titania.cobuild.collins.co.uk/boe_info.html
British National Corpus: http://info.ox.ac.uk/bnc
Dictionary of American Regional English: http://polyglot.lss.wisc.edu/dare/dare.html