2

A Linguistic 'Time Capsule': The Newcastle Electronic Corpus of Tyneside English

Will Allen, Joan Beal, Karen Corrigan, Warren Maguire and Hermann Moisl[1]

1 Introduction

The general goal of this chapter is to outline the models and methods underpinning the Newcastle Electronic Corpus of Tyneside English (NECTE), created by the amalgamation of two separate corpora of recorded speech from the same geographical location. The earliest of these was collected in the late 1960s and early 1970s as part of the Tyneside Linguistic Survey (TLS) funded by the Social Science Research Council (SSRC) (see Strang, 1968; Pellowe *et al*., 1972; Pellowe and Jones, 1978; and Jones-Sargent, 1983). The more recent of the two was created between 1991 and 1994 for a project entitled Phonological Variation and Change in Contemporary Spoken English (PVC), which was supported by the Economic and Social Research Council (ESRC) (see Milroy *et al*., 1997). More specifically, the chapter addresses four topics: (i) the objectives of the NECTE enhancement programme and the original aims of the TLS and PVC projects that are its foundation; (ii) the initial state of the sources on which the NECTE corpus is built; (iii) procedures for the amalgamation of these sources; and (iv) projected further developments of the resultant corpus and preliminary linguistic analyses of it.

2 NECTE aims and objectives

In 2001, the NECTE project was funded by the AHRB with the aim of providing an enhanced electronic corpus resource. It was to be Text Encoding Initiative (TEI)-conformant and would eventually be made

available to the public and to the research community in a variety of formats: digitized sound, phonetic transcription, orthographic transcription and grammatical markup, all aligned and downloadable from the Web.

2.1 TLS aims and objectives

The chief aim of the TLS was to determine the 'ecology' of urban varieties of English (that is, what kinds of variation exist), using a radical and rigorous statistical methodology that had evolved in opposition to the already predominant Labovian paradigm (see Labov, 1972, and Trudgill, 1974, for instance). Rather than pre-selecting 'salient', linguistic variables and correlating these with a narrow range of external indices, such as social class, the TLS grouped speakers and analysed their similarity to one another by comparing their data sets across a multitude of variables simultaneously. Each informant would thus be assigned a unique position in linguistic 'space', and differences between speakers would be evident in the manner in which these clustered relative to one another. 'Linguistic' clusters (grammatical, phonological and prosodic variants) could then be mapped onto 'social' clusters, likewise arrived at by multivariate analyses of the subjects' scores on a wide range of social and lifestyle factors from 'educational level' to 'commitment to taste in décor'.[2]

While the theoretical approach, methodology and initial outcomes from the TLS aroused a certain amount of interest, the project was perceived to be overly complex at the time, as Milroy (1984, p. 207) articulates in her statement:

> Although many would feel sympathetic to the aims of this ambitious project, the very punctiliousness of the Tyneside Linguistic Survey researchers has led to an imbalance in favour of methodology and theory and a relative weakness on results.

It is unsurprising, therefore, that despite some 'stimulating and innovatory' public outputs (Milroy, 1984, p. 207), the research programme was never fully completed and, indeed, remained largely forgotten until the archiving and transcription projects of Beal and Corrigan between 1994 and 2001 (see Beal (1994–5), Beal and Corrigan (1999), Beal and Corrigan (2000–2001), Beal, Corrigan, Duchat, Fryd and Gerard (1999–2000), Corrigan (1999–2000), and also section 3.1 below).[3] Thus, as well as preserving and disseminating what has become a valuable historical record of Tyneside English, the NECTE project brings closure on

the one hand and a new beginning on the other to an important, but neglected, chapter in the history of sociolinguistics.

2.2 PVC aims and objectives

As noted in section 1 above, we have also incorporated and enhanced spoken data collected under the auspices of the PVC project. In contrast to the TLS, this latter research produced substantial outputs, which together have made a very significant contribution both to the methodology of sociophonetics and to our understanding of the nature of dialect levelling in late twentieth-century Britain (see, for instance, Milroy *et al*., 1997; Docherty and Foulkes, 1999; Watt and Milroy, 1999; Watt, 2002). However, although the focus of research by the PVC team was on *phonological* variation and change, the data hold a great deal of valuable and exploitable analytical information for other fields of research. In our efforts to encourage more divergent linguistic investigations of this resource, the NECTE project has, therefore, made the PVC corpus available to a wider range of end-users than those envisaged when the interviews were originally conducted. We believe that doing so is crucial, given the richness of the data for morphosyntactic studies, for instance, as has already been demonstrated in Beal and Corrigan (2002, 2005a, 2005b) and Beal (2004b).

Above all, the amalgamation of these two data sets incorporating Tyneside speakers from different age, class and sex groupings between the middle and end of the twentieth century makes the NECTE corpus invaluable for both real- and apparent-time studies of internal and external variation on a number of linguistic levels, as argued in Beal and Corrigan (2000a, 2000b, 2000c).[4]

3 The sources on which the NECTE corpus is built

3.1 The Tyneside Linguistic Survey (TLS)

To judge from the unpublished papers and public output of the TLS, its main aim, as described in section 2.1, was to determine the nature and extent of linguistic variation among Tynesiders and how this might be correlated with a range of social and lifestyle factors. To realize this research aim, the TLS team created a corpus of materials relating to Tyneside English consisting of the following components:

- A collection of audio-taped interviews with speakers who were encouraged to talk about their life histories and their attitudes to the local dialect. In addition, at the end of each interview, infor-

mants were asked for acceptability judgements on constructions containing vernacular morphosyntax, and whether they knew or used a range of traditional dialect words. Interviews varied somewhat in length but lasted 30 minutes on average, and were recorded onto analogue reel-to-reel tape, the standard audio-recording technology of the time.

- Detailed social data for each speaker.
- Orthographic and phonetic transcriptions recorded onto index cards. Approximately 200 of these were completed for each interview, equating to the first ten minutes or so of the recording session. Index card transcriptions conveyed interviewee turns only and each brief section of audio was annotated for the following types of information: (i) Standard English orthography; (ii) a corresponding phonetic transcription of the audio segment; and (iii) some associated grammatical, phonological and prosodic details (see Figure 2.4 below).
- Digital electronic text files containing encoded versions of the phonetic transcriptions (1-Alpha codes) as well as separate ciphers conveying grammatical, phonological and prosodic (2-Alpha and 3-Alpha) information (see Figure 2.3 below).
- Digital electronic text files including additional codes that conveyed different kinds of social data for each speaker.[5]

Following the end of the SSRC award, the audio tapes and index card sets were stored in the Department of English Language (now part of the School of English Literature, Language and Linguistics) at the University of Newcastle upon Tyne. In addition, John Local, one of the TLS researchers, deposited six audio recordings with the British Library Sound Archive. The electronic files, which had been crucial to implementing the unique variationist methodology of the TLS project, were lodged with the Oxford Text Archive (OTA).

In 1994–95, Joan Beal at Newcastle University secured funding from the Catherine Cookson Foundation to: (i) salvage the rapidly deteriorating reel-to-reel audio tapes by re-recording them onto cassette tape, (ii) catalogue them alongside the social data, and (iii) archive the tapes, the index card sets and documentation associated with the TLS project in a new Catherine Cookson Archive of Tyneside and Northumbrian Dialect at the University of Newcastle. Had it not been for this funding, these 'hard-won sounds' of mid-twentieth-century Tyneside would, as Widdowson (2003, p. 84) puts it, have simply become dispersed 'particles of ferric oxide'.

Since 2001, the NECTE project has based the TLS component of its enhancement scheme on the material in the Catherine Cookson Archive and the British Library Sound Archive and on the electronic holdings of the digital files at the OTA. As restoration and digitization efforts progressed, it became evident that only a fragment of the projected TLS corpus had survived. Unfortunately, it still remains unclear exactly how much material has, in fact, been permanently lost. The crux of the matter is that the information in unpublished TLS project documentation (as well as that in the public domain) does not allow one to decide with any certainty how large the corpus originally was. We are not sure, for example, how many interviews were conducted, and the literature gives conflicting reports. Pellowe *et al*. (1972, p. 24), for example, claim that there were 150, whereas Jones-Sargent (1983, p. 2) mentions the higher figure of 200. It is also unknown how many of the original interviews were orthographically and phonetically transcribed. Jones-Sargent (1983) used 52 (digitally encoded) phonetic transcriptions in her computational analysis, but the TLS material includes seven electronic files that we recovered from the OTA, but that she did not use. As such, there were clearly more than 52 phonetic transcriptions, but was the ultimate figure 59, or were further files digitized but never passed to the OTA?

All one can reasonably do in this situation, therefore, is catalogue and enhance what currently exists and, to date, NECTE has been able to identify 114 interviews, but not all corpus components survive for each. Specifically, there remain:

- 85 audio recordings, of which three are badly damaged (110, 111 and 113 in Table A2.1 of the Appendix). For the remaining 29 interviews, the corresponding analogue tape is either blank or simply missing
- 57 index card sets, all of which are complete
- 61 digital phonetic transcription files
- 64 digital social data files.

The distribution of these materials across interviews is shown with an 'X' in Table 2.A1 of the Appendix. There is no natural order to the interviews as such, so they are arranged there in descending order depending on how many corpus components they still retain. Those interviews that have all four components are at the top of the table, followed by others with only three components, and so on. When the interviews are arranged in this way, it is easily seen that, out of 114 interviews, only 1–37 are complete in the sense that an intact audio

recording, an index card set, and electronic phonetic transcription and social data file exist, all the others (38 onwards) being fragmentary to greater or lesser degrees.

3.2 The Phonological Variation and Change in Contemporary Spoken English (PVC) corpus

The primary PVC materials comprise 18 digital audio-taped interviews, each of up to 60 minutes' duration. Self-selected dyads of friends or relatives conversed freely about a wide range of topics with minimal interference from the fieldworker. As such, the interview format was that which, in the terms of Macaulay (2005, pp. 21–2), 'lies between the monologues of individual interviews and the polyphony of group sessions'. It has the advantage of generating 'unstructured conversations in optimal recording situations' which also delimit the observer's paradox and the concomitant possibility of accommodation.

In contrast to the detailed phonetic transcriptions provided by the TLS, the PVC team restricted their transcription to those specific lexical items in phonetic context that they were interested in analysing from auditory and/or acoustic perspectives. No systematic orthographic transcription of the material, such as that produced by the TLS, was ever attempted. As Table 2.1 (adapted from Watt and Milroy, 1999, p. 27) demonstrates, the PVC project did, however, record some social data, though it was not as detailed as that of the TLS team, since they restricted their categorization of subjects to age, gender and 'broadly defined socio-economic class' (Watt and Milroy, 1999, p. 27).

4 Procedures for the amalgamation of these sources

One of the most complex aspects of the NECTE encoding project is that the resultant digital corpus aimed to provide four different levels of data representation, namely, (i) audio, (ii) orthographic transcription, (iii) grammatical markup, and (iv) phonetic transcription. In what follows below, we provide a description of the methods used to create

Table 2.1 Design of PVC fieldwork sample

Working class (WC)				**Middle class (MC)**			
Younger (15–27)		**Older (45–67)**		**Younger (15–27)**		**Older (45–67)**	
Male	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
4	4	4	4	4	4	4	4

these digital representations and a discussion of the most significant problems encountered when devising or recreating each level.

4.1 The audio data

Both the TLS and PVC corpora are preserved on audio tape and, unsurprisingly therefore, we view the primary NECTE data representation as being audio. The relative 'youth' and high sound quality of the PVC recordings has enabled a largely trouble-free, though still fairly time-consuming, preparation of the data for the purposes of the NECTE project. The TLS recordings, on the other hand, have been rather more problematic, since they required a considerable degree of restoration. The original analogue recordings, both reel-to-reel and the cassette versions, which came about as a result of their 'rescue' in 1994–95 (as outlined in section 3.1 above), were first digitized at a high sampling rate. All the TLS recordings included in NECTE were digitized from the cassette versions in WAV format at 12000 Hz 16-bit mono and were enhanced by amplitude adjustment, graphic equalization, clip and hiss elimination, as well as regularization of speed.[6] All PVC recordings were digitized in WAV format direct from the original DAT tapes and required no additional adjustment.

4.2 Orthographic transcription

The audio content of the TLS and PVC corpora has been transcribed orthographically and this is also included as a level of representation in the NECTE corpus. As noted in Preston (1985), Macaulay (1991), Kirk (1997) and, more recently, in Cameron (2001) and Beal (2005), representing vernacular Englishes orthographically, by using 'eye-dialect', for example, can be problematic on a number of different levels. Attempting to convey 'speech realism', in Kirk's (1997, p. 198) sense, can lead, for example, to an unwelcome association with negative racial or social connotations and there are theoretical objections too in that devising non-standard spellings to represent certain groups of vernacular speakers can make their speech appear more differentiated from mainstream colloquial varieties than is actually warranted. With these caveats in mind, we outline below the 'trade-offs' (see Tagliamonte, Volume 1) adopted for the NECTE project in this regard. Two issues, in particular, have exercised us in our attempt to transcribe the audio data orthographically in a maximally efficient and accurate manner that simultaneously encodes the nuances particular to spoken Tyneside English which a range of end-users might conceivably want annotated for them. The first of these relates to the application of the

conventional spellings associated with standard written British English to a non-standard spoken dialect. Leaving aside, for present purposes, the question of writing as opposed to speaking conventions, Tyneside English speech differs significantly from standard spoken English across all linguistic levels, from phonetic to pragmatic. As such, it would have been uneconomic in the extreme to attempt to render all of these potential differences in the NECTE Orthographic Transcription Protocol (OTP). This seemed particularly justified with respect to the phonetic level, since it was always our intention to provide International Phonetic Alphabet (IPA) transcriptions for a carefully selected cohort of the PVC corpus sound files and because the TLS project had already bequeathed to us a highly detailed phonetic transcription of much of their material which we planned to authentically reformulate (see section 4.4 below). As such, no attempt was made to represent the non-standard phonology of Tyneside English with semi-phonetic spelling, for example. Hence, we have chosen to ignore popular representations such as the distinctive Tyneside pronunciation /na:/ as <knaa> for Standard English <know>. However, in cases where local vernacular renditions are either lexically or morphologically distinct from standard British norms, a representation was agreed by way of an OTP (see Poplack, 1989; and Lawrence *et al*., 2003) and adhered to consistently. For instance, if a particular lexeme had an established tradition, having been recorded in a published dialect glossary such as Brockett (1825), Heslop (1892–94), Geeson (1969), Dobson (1974), Graham (1979), Griffiths (1999), Douglas (2001) or Moody (forthcoming), then this spelling was adopted. A lexicon of dialectal lexical items was compiled and added to the OTP as transcription proceeded, with cross-references, where appropriate, to the established glossaries, as in the examples from a range of semantic fields given in Table 2.2.

Table 2.2 Orthographic representation of dialectal lexical items

Gloss	**Moody** (forthcoming) **spelling**	**NECTE protocol spelling**
'food, packed lunch'	*Bait* (p. 38)	*bait*
modifier, e.g. 'canny few', meaning 'a lot'	*Canny* (p. 112)	*canny*
'sticky/slimy'	*Claggy* (p. 132)	*claggy*
'street chasing game'	*Kick-the-block* (p. 345)	*kick the block*
'sledge-hammer'	*Mell* (p. 395)	*mell*
'birds, especially sparrows'	*Spuggies* (p. 535)	*spuggies*

As Poplack (1989) rightly points out, any large-scale textual transcription exercise is subject to human error of various sorts. Our particular problems in this regard are exacerbated by the fact that the TLS tapes are now several decades old and, as noted already, have become degraded in various ways, so that it is often difficult or even impossible to hear what is being said. Acoustic filtering in the course of digitization, such as that described immediately above, improved audibility in some, but by no means all, cases. To offset these difficulties, we have availed ourselves of certain orthographic transcriptions made by the TLS project team back in the 1960s and 1970s when the original tapes were still in good condition, but, as Table A1 reveals, these cover only part of the corpus.[7] To ensure accuracy, therefore, we found it necessary to conduct four transcription passes through the audio files. The first of these established a base text, the second and third were correction passes to improve transcription accuracy, and the fourth established uniformity of transcription practice across the entire corpus.

4.3 Part-of-speech tagged orthographic transcription

Grammatical tagging was crucial to the NECTE programme as a level of data representation. The annotation scheme chosen was determined by what was possible within the timescale of the project, subject to the following constraints:

- Existing tagging software had to be used.
- The tools in question had to encode non-standard English reliably, that is, without the need for considerable human intervention in the tagging process and/or for extensive subsequent proofreading.

Having reviewed the full range of tagging software currently available and with these constraints in mind, the CLAWS tagger, developed for annotating the BNC by UCREL (University Centre for Computer Corpus Research on Language) at Lancaster University, UK, was selected. It fulfils NECTE's requirements in that it is a mature system developed over many years, which has consistently achieved an accuracy rate of 96–97 per cent in relation to the BNC corpus. The NECTE (that is, not the TLS) orthographic transcriptions of the TLS and the PVC audio were part-of-speech tagged by the CLAWS4 tagger using the UCREL C8 tagset and Figure 2.1 contains a sample of the resulting tagged output for the sentence: *and eh I lived in with my mother for not quite two year but varnigh*.[8]

```
<u who="informantTL Sg37">
<w type="CC" lemma="and">and</w>
<w type="UH" lemma="eh">eh</w>
<w type="PPIS1" lemma="i">i</w>
<w type="PPIS1" lemma="i">i</w>
<w type="VVD" lemma="live">lived</w>
<w type="RP" lemma="in">in</w>
<w type="IW" lemma="with">with</w>
<w type="APPGE" lemma="my">my</w>
<w type="NN1" lemma="mother">mother</w>
<w type="IF" lemma="for">for</w>
<w type="XX" lemma="not">not</w>
<w type="RG" lemma="quite">quite</w>
<w type="MC" lemma="two">two</w>
<w type="NNT1" lemma="year">year</w>
<w type="CCB" lemma="but">but</w>
<w type="VV0" lemma="varnigh">varnigh</w>
</u>
```

Figure 2.1 CLAWS output

4.4 Phonetic transcription

NECTE includes partial phonetic transcriptions of the TLS and PVC interviews. The TLS phonetic transcriptions require some detailed discussion, so they will be treated in section 4.4.2 below after a brief description of the phonetic transcription practices used for the PVC data.

4.4.1 PVC phonetic transcription

Sample phonetic transcriptions of the PVC materials are provided for comparison with the TLS transcriptions. These are far less extensive than TLS on account of the extremely time-consuming nature of the process (as articulated in section 4.4.2 below). Previous research, such as Kerswill and Wright (1990), as well as consultation with sociophoneticians (Gerard Docherty, Paul Foulkes, Paul Kerswill and Dom Watt) with expertise in north-eastern dialects and other potential end-users, confirmed that most researchers whose primary interest was in phonetics would prefer to do their own analyses, so a decision was taken to provide only broad transcriptions of a stratified subsample. The first five minutes of each of six PVC tapes was transcribed, giving samples of twelve speakers in all. This was done at a 'broad' phonetic level.

4.4.2 TLS phonetic transcription

To realize its main research aim stated in section 2 above, the TLS had to compare the audio interviews it had collected at the phonetic level of representation. This required the analogue speech signal to be discretized into phonetic segment sequences, in other words to be phonetically transcribed. The standard method is to select a transcription scheme, that is, a set of symbols each of which represents a single phonetic segment (for example, that of the IPA), and then to partition the linguistically relevant parts of the analogue audio stream such that each partition is assigned a phonetic symbol. The result is a set of symbol strings each of which represents the corresponding interview phonetically. These strings can then be compared and, if also given a digital electronic representation, the comparison can be done computationally.

The TLS team generated phonetic transcriptions of a substantial part of its audio materials, and they are included in the NECTE corpus. However, in order to make them usable in the NECTE context they have required extensive restoration. The sections below describe the TLS phonetic transcription scheme (4.4.2.1) and the rationale for and the restoration of the TLS electronic phonetic files (4.4.2.2).

4.4.2.1 TLS phonetic transcription and digital encoding schemes The TLS made the simple, purely sequential transcription procedure described above the basis for a rather complex hierarchical scheme that would eventually represent the phonetics and phonology of its corpus.

That scheme has to be understood if its phonetic data are to be competently interpreted, and it is consequently explained in detail below.

The TLS team developed its hierarchical phonetic transcription scheme in order to capture as much of the phonetic variability in the interviews as possible. To see exactly how this might be achieved, consider what happens when data generated by a sequential transcription procedure are analysed and, more specifically, the transcribed interviews are compared. An obvious way to do the comparison is to count, for each interview, the number of times each of the phonetic symbols in the transcription scheme being used occurs. This process yields a phonetic frequency profile for each of the interviews, and the resulting profiles can then be compared using a wide variety of methods. Unfortunately, such profiles fail to take into account a commonplace of variation between and among individual speakers and speaker groups, namely that different speakers and groups typically distribute the phonetics of their speech differently in distinctive lexical environments. Frequency profiles of the sort in question here only say how many times each of the various speakers uses phonetic segment *x* without regard to the possibility that they distribute *x* differently over their lexical repertoires. The hierarchical TLS transcription scheme was designed to capture such distributional variation.

The scheme is similar to the manner of specifying the lexical distribution of vowels in any given English accent used by Wells (1982), whereby comparisons are made using a number of standard 'lexical sets'. As the name implies, the latter is a set of words which can be related in some respect. Those described by Wells (1982) define sets of words which, taking RP and General American as reference points, have shared phonological histories, and give extensional definitions of the phonemes of those accents. Hence, the KIT set {*ship, rib, dim, milk, slither, myth, pretty, build, women, busy* …}, for example, defines the phoneme /I/ in Standard American and Received Pronunciation British English. In a similar way, the TLS used the phonemes of RP (only) as a basis for the definition of lexical sets.

The TLS hierarchical transcription scheme, which exploits a similar method, has three levels:

- The top level, designated 'Overall Unit' (OU) level, is a set of lexical sets where OU = {{ls_1}, {ls_2}…{ls_m}}, such that each {ls_i} for $1 < i < m$ extensionally defines one phoneme in RP, and *m* is the number of phonemes in RP. The purpose of this level was to provide a standard

relative to which the lexical distribution of Tyneside phonetic variation could be characterized.

- The bottom level, designated 'State', is a set of phonetic symbol sets where State = $\{\{ps_1\}, \{ps_2\}...\{ps_m\}\}$. There is a one-to-one correspondence of lexical sets at the OU level and phonetic symbol sets at the State level such that the symbols in $\{ps_i\}$, for $1 < i < m$, denote the phonetic segments that realize the OU $\{ls_i\}$ in the fragment of Tyneside English that the TLS corpus contains.
- The intermediate Putative Disystemic Variable (PDV) level proposes (thus 'putative') groupings of the phonetic symbols in a given State set $\{ps_i\}$ based (as far as the existing TLS documentation allows one to judge) on the project's perceptions of the relatedness of the phonetic segments that the symbols denote. These PDV groups represent the phonetic realizations of their superordinate OUs in a less fine-grained way than the State phonetic symbol sets do. A detailed example of the scheme taken from Jones-Sargent (1983, p. 295) is given in Figure 2.2.

OU	PDV	(code)	states	lexical examples
1 [i:]	i:	0002	i [i] [i] [i] [i] [i]	week, treat, see
	I	0004	[i] [i] [i] [i] [i]	week, relief
	ɛ	0006	[e] [ɛ] [e] [ɛ]	beat
	eI	0008	[ɛi] [əi] [ɛ] [ɛi]	see
	Iə	0010	[iɛ] [iɛ] [iə]	feed
	Ii	0012	ii(back) ii(low) [i]	we, see
2 [I]	I	0014	[i] [i] [i] [i] [i]	fit, big, till
	ə	0016	[ə] [ə] [ə] [ə] [ɜ] [ɜ]	shilling
	Iə	0018	[iə] [iə] [iə]	did
	ɜ:	0020	[ɜ] [ɜ] [ɛ]	shilling
	ɛə	0022	[ɛ] [ɛ]	miss, big

Figure 2.2 TLS coding scheme for realizations of the OUs [i:] and [I]

The OU i: defined by the lexical set (from which there are examples in the rightmost column of Figure 2.2) can be realized by the phonetic segment symbols in the States column, and these symbols are grouped by phonetic and lexical relatedness in the PDV column. This transcription scheme captures the required distributional phonetic information by allowing any given State segment to realize more than one OU. Note that several of the State symbols for OU i: occur also in the OU I. What this means is that, in the TLS transcription scheme, a State phonetic segment symbol represents not a distribution-independent sound, but a sound in relation to the phonemes over which it is distributed.

The implications of this can be seen in the encoding scheme that the TLS developed for its transcription protocol so that its phonetic data could be computationally analysed. Each State symbol is encoded as a five-digit integer. The first four digits of any given State symbol designate the PDV to which the symbol belongs, and the fifth digit indexes the specific State within that PDV. Thus, for the OU i: there are six PDVs, each of which is assigned a unique four-digit code.[9] For a given PDV within the i: OU, say I, the first of the state symbols in left-to-right order is encoded as 00041, the second as 00042, and so on. Now, note that the State symbols 00023 and 00141 are identical, that is, they denote the same sound. Crucially, however, they have different codes because they realize different phonemes relative to OU, or, in other words, the different codes represent the phonemic distribution of the single sound that both the codes denote.[10]

4.4.2.2 Restoration of the TLS phonetic transcriptions The phonetic transcriptions of the TLS interviews survive in two forms, that is, as a collection of index cards and as electronic files. Each electronic file is a sequence of the five-digit codes just described, a random excerpt from one of these files being given in Figure 2.3.

02441 02301 02621 02363 02741 02881 02301 01123 00906 02081-&&&& *

02322 02741 02201 02383 02801 02421 02501 01443 01284 00421 02021 00342

02642 02164 02721 02741 04321-&&&&

02621 02825 02301 02721 02341 02642 02541 00503 00161 00246 12601 01284

02781 02561 02363 02561 02881 07641 02941-&&&&

* The sequences designated by the *TLS* coders as '-&&&& are end-of-line markers.

Figure 2.3 A sample of a TLS electronic file of five-digit codes

Initially, from NECTE's perspective, these electronic files appeared to be a labour- and time-saving alternative to keying in the numerical codes from the index cards. However, a peculiarity that stems from the original electronic data entry system used by the computing staff who had been entrusted with the task of creating the files from the TLS team's original index cards meant that they had to be extensively edited by NECTE personnel when the files were returned to us from the OTA. The problem arose from the way in which the five-digit codes were laid out by the TLS researchers on the index cards, as in Figure 2.4.

For reasons that are no longer clear, all the consonant codes (beginning (0220(1)) in line 4 of Figure 2.4) were written on one line, and all of the vowel codes appear on the line below ((0062(7)) of line 5 in Figure 2.4). When the TLS gave these index cards to the University of Newcastle data entry service, the typists entered the codes line by line, with the result that, in any given electronic line, all the consonant codes come first, followed by the vowel codes. This difficulty pervades the TLS electronic phonetic transcription files. While it had no impact on the output of the TLS team (given that they were examining codes in isolation and since phonetic environment was already captured by their hierarchical scheme), it was highly problematic for the NECTE enhancement of the original materials.

Simply to keep this ordering would have made the phonetic representation difficult to relate to the other types of representation planned for the NECTE enhancement scheme. The TLS files were therefore edited with reference to the index cards so as to restore the correct code sequencing, and the result was proofread for accuracy.[11]

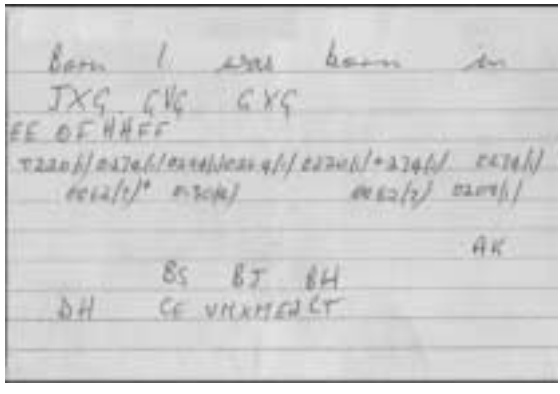

Figure 2.4 A sample of a TLS index card

The only exception to this restoration procedure is the files for the Newcastle speakers. Because neither the audio recordings nor the index card sets for these speakers survive, restoration of the correct sequencing would have been a hugely time-consuming task, and one that could not be undertaken within the limited time available to the NECTE project. Even in their unordered state, however, these files are still usable for certain types of phonetic analysis such as those that involve segment frequency counts, and they are included in NECTE in their present state for that reason. Moreover, the formatting of numerical codes in these files differs from that in the other TLS-based files, where the codes are in a continuous sequence. For the Newcastle files, the original TLS formatting has been retained: the numerical codes are arranged in a sequence of code strings each of which is terminated by a line break, where a code string in the sequence corresponds to a single informant utterance. The motivation was to facilitate reordering of the codes if this is ever undertaken in future (if, for example, the audio files or the index card sets for the Newcastle group should ever come to light).

4.5 Content alignment

The NECTE project felt that the usefulness of its corpus would be considerably enhanced by the provision of an alignment mechanism relating the representational types described in sections 4.1–4.4 above to one another, so that corresponding segments in the various layers can be conveniently identified and simultaneously displayed. The decision to provide such an instrument immediately raises the question of granularity: how large should the alignment segments be? Should the representational types be aligned at the level of the phonetic segment or would it be more appropriate to set the alignment at a lower level of granularity such as sentence or even utterance? In addition to considering the divergent discourse dynamics of the TLS and PVC corpora, our evaluation of this crux also had to take into account research utility on the one hand, and feasibility in terms of project cost on the other, that is, would word-by-word alignment, say, be useful enough from the perspective of potential research on the corpus to justify the considerable effort required to manually insert the numerous markers necessary for so fine-grained a resolution?

For the TLS materials, the format of the interviews made alignment at the granularity of utterance the natural choice. This is because a typical interview consists of a succession of interviewer–question, interviewee–answer pairs in which the utterance boundaries are generally

clear-cut (as is the norm for adjacency pairs more generally, according to Sacks *et al*., 1974). There is some degree of overlap on account of interruption and third-party intervention, but this is infrequent enough to be handled fairly straightforwardly within an utterance-aligned framework.

The PVC materials, however, presented a rather different discourse situation since the interviews were considerably more loosely structured. In the first place, the interviewer's role is to monitor rather than participate, making their contribution almost negligible. Second, the setting is designed such that a majority of the conversation emanates from pairs of subjects who are either friends or relatives and therefore knew each other very well in advance of the interview. Unsurprisingly, therefore, overlaps in this case are the norm rather than the exception. Attempting to disentangle the speakers would, on the one hand, require very detailed markup (with consequent additional project costs), and, on the other, would necessitate ad hoc decisions about conversational structure, thereby imposing an undesirable pre-analysis on the data (see Stenstrom and Svartvik, 1994). Assuming the need for a uniform alignment mechanism across the entire corpus, it was clear that alignment at the utterance level was, therefore, impractical.

What were the alternatives? More detailed alignment at the granularity of the phonetic segment or of the word was ruled out on account of excessive manual effort. So too was alignment on the basis of syntactic unit, since this would have necessitated either manual syntactic markup (which would once again have been expensive) or access to a reliable automatic parser for the highly vernacular English that the corpus contains, which, to our knowledge, does not yet exist. The one choice remaining, which seemed both appropriate to the distinctive discourses of the TLS and PVC and was feasible in terms of cost, appeared to be alignment by 'real-time' interval, which was therefore the method that the NECTE team eventually adopted.

Our real-time interval alignment mechanism works as follows. It begins with the observation that real time (that is, time as it is conceived by humans in day-to-day life) is meaningful only for the audio level of representation in the corpus. By contrast, text, be it orthographic, tagged, or a sequence of phonetic symbols, has no temporal dimension. A time interval t is therefore selected and the audio level is partitioned into some number n of length-t audio segments s: $s(t \times 1)$, $s(t \times 2)$ … $s(t \times n)$, where '×' denotes multiplication. Corresponding markers are then inserted into the other levels of representation such that they demarcate substrings corresponding to the audio segments.

Hence, for the audio segment $s(t \times i)$, for some i in the range $1...n$, there are markers in the other representational levels which identify the corresponding orthographic, phonetic and part-of-speech tagged segments. In this way, selection of any segment s in any level of representation allows the segments corresponding to s in all the other levels to be identified.

A time interval of 20 seconds was selected for this procedure on the grounds of project cost and end-user need. With regard to the former, it is obvious that the shorter the interval, the greater the effort of marker insertion. The increase is more than linear as the interval shrinks, that is, markup for a one-second interval takes more than 20 times longer than markup for a 20-second interval, due to the simple mechanics of starting and stopping the audio stream in exactly the right place and then deciding where to put the markers in the other levels. A 20-second interval was thus felt to be a cost-effective choice that also coincided with the potential demand for usability. With respect to the latter, 20-second chunks were found to yield about the right amount of aligned text from the various levels of representation on a typical computer screen when all the levels are simultaneously displayed, as in Figure 2.5 in the next section.[12]

4.6 Document structuring

NECTE is encoded using Text Encoding Initiative (TEI)-conformant Extensible Markup Language (XML) syntax. XML (http://www.w3.org/XML/) aims to encourage the creation of information resources that are independent both of the specific characteristics of the computer platforms on which they reside (Macintosh versus Windows, for example), and of the software applications used to interpret them. To this end, XML provides a standard for structuring documents and document collections. TEI defines an extensive range of XML constructs as a standard for the creation of textual corpora in particular. Together, these are emerging as world standards for the encoding of digital information, and it is for this reason that NECTE adopted them.

Specifically, the NECTE corpus is a TEI-conformant XML document in the TEI local processing format sense, as specified in the *Guidelines for Text Encoding and Interchange* (Sperberg-McQueen & Burnard, 2002, ch. 28).[13] To be TEI-conformant, an XML document has to be validated relative to the TEI Document Type Definition (DTD). NECTE's selection of a validator was based on information provided by Thijs van den Broek's technical report *Benchmarking XML-editors* (2004), a version of which is available on the Arts and Humanities Data Service (AHDS)

website (http://ahds.ac.uk/creating/information-papers/xml-editors/). We chose the oXygen XML Editor (http://www.oxygenxml.com/) since it provides facilities not only for the creation of XML documents but also for their validation in relation to user-defined DTDs. The NECTE corpus document has, therefore, been validated relative to the TEI DTD by oXygen.[14]

To make all of this more concrete, consider the following severely truncated excerpt from the actual NECTE corpus in Figure 2.5.

```
<teiCorpus.2>
<teiHeadertype='corpus'>
<fileDesc>…</fileDesc>
<encodingDesc>…</encodingDesc>
<profileDesc>…</profileDesc>
<revisionDesc>…</revisionDesc>
</teiheader>
<TEI.2 id="TLSgoi">
<teiHeadertype="text">
<!--Header information-->
</teiHeader>
<text>
<group>
<text id='TLSg01audio'>
<body>
<audio entity='TLSaudiog37'>
<body>
</text>
<text id='TLSg01NECTEortho'>
<body>
<u who="informantTLSg37"><anchor id="TLSg37NECTEortho0000"/>tIs what's that </u><u
who="interviewerTLSg37">g</u><u who="informantTLSg37">e<pause/>five two</u><u who"interviewerTLSg37">
thanks<pause/>ta<pause/> eh could you tell us first eh where you were born please<event desc="interruption"/>
<unclear/></u><u who="informantTLSg37"> i was born at eleven victoria street <pause/> gateshead </u><u
who="interviewerTLSg37">eh<pause/>aye yeah whereabouts is that again…
</body>
</text>
<text id='TLSg01TLSortho'>
<body>
```

Figure 2.5 Truncated excerpt from the XML version of NECTE

```
<u who="informantTLSg37"><anchor id="TLSg37TLSortho0000"/> I was born at eleven Victoria Street Gateshead thats
just against the <pause/> the flats <anchor id="TLSg37TLSortho0020"/> you know no Barney Close the old <pause/>
Victoria Street eleven Victoria Street oh well I my mother got out they were building houses for the people then down the
Old Fold and I went down the Old <anchor id="TLSg37TLSortho0040"/> Fold to live
</body>
</text>
<text id='TLSg01phonetic'>
<body>
<u who="informantTLSg37"><anchor id="TLSg37phonetic0000"/> 01304 02941 02641 02201 00626 02741 08760 02301
02081 02781 00244 02561 02021 02741 02561 00144 02421 02263 00626 02861 17801 02621 02262 02861 00023 02301
02442 01123 02301 02623 02365 02603 00342 02301 09040 02521 00823 02623 02442 11202 02741 02623 09030 08440
08580 02603 02541 02801 00342 02301 28803 <anchor id="TLSg37phonetic0020"/> 02741..
</body>
</text>
<text id='TLSg01tagged'>
<body>
<u who="informantTLSg37"><anchor id="0000"/> <s> <w type="VVN" lemma="see"> seen </w> <w
type="II" lemma="at"> at </w> <w type="AT" lemma="the"> the </w> <w type="NN2" lemma="picture"> pictures </w>
<w type="VBDZ" lemma="be"> was </w> <w type="UH" lemma="ehm"> ehm </w> <w type="RR" lemma="so"> so
</w> <w type="PPIS1" lemma="i"> i </w> <w type="VVD" lemma="marry"> married </w> <w type="AT1" lemma="an">
an </w> <w type="NN1" lemma="axe"> axe </w> <w type="NN1" lemma="murderer"> murderer </w>..
</body>
</text>
</group>
</text>
</TEI.2>
</teiCorpus.2>
```

Figure 2.5 Continued

As can be seen, the textual content is surrounded by and interspersed with a multitude of tags enclosed by angle brackets. These serve to specify the many features of the NECTE corpus structure: <u who="informantTLSg37">, for instance, identifies a particular TLS informant and indicates that the speaker turn is about to begin; <w type="NN2" lemma="picture"> pictures </w> identifies 'pictures' as a word (<w>) of lexical type NN2 and lemma 'picture', and so on.

5 Projected further developments of NECTE and preliminary linguistic analyses

5.1 Further enhancement plans

Two developments of NECTE are currently projected in the short to medium term.

5.1.1 Provision of visualization and transformation facilities

Adoption of TEI-conformant XML requires no justification in principle as it is a world standard, but it can be an obstacle to users of the corpus in practice. The documentation page of the project website (http://www.ncl.ac.uk/necte/documentation.htm) observes, rather economically, that 'familiarity with XML and TEI is assumed throughout'. Users not familiar with these standards may find the pervasive markup tags in the NECTE files a distracting encumbrance and yearn for the good old days of plain text files. This is not an unreasonable position. XML was never intended to be reader-friendly. It is a markup language that provides a standard for the structuring of documents and document collections, and, though XML-encoded documents are plain text files that *can* be read by humans, in general they should not be. For an XML document to be readily legible, software that can represent the structural markup in a visually accessible way is required. For example, XSLT (Extensible Stylesheet Language Transformations) (http://www.w3.org/TR/xslt) can transform an XML-encoded document into an HTML-encoded one that can then be viewed using any standard web browser. Similarly, search and analysis of NECTE or any other XML-encoded corpus requires software to interpret the markup.

XML-aware software visualization and analysis tools are gradually becoming available. The Oxford University Computing Service's Xaira system (http://www.oucs.ox.ac.uk/rts/xaira/), for instance, is 'a general purpose XML search engine, which will operate on any corpus of well-formed XML documents. It is, however, best used with TEI-conformant documents'.[15] For user convenience, therefore, in further developing the NECTE corpus, our next priority is to provide XSLT style sheets that generate, on the one hand, HTML versions of the corpus content for accessible visualization using standard web browsers, and, on the other, plain text versions that can be used with existing applications that are not XML-aware.

5.1.2 Addition of materials

The catalogue of TLS materials described in section 3.1 above and outlined in Table 2A.1 of the Appendix shows that, in addition to the core

of complete interviews that NECTE now contains, there is a penumbra of fragments, the inclusion of which could usefully augment the corpus in various ways (though, of course, they can never be subject to an identical alignment procedure since certain levels of representation are unavailable). Hence, interviews 38 and 39 can be included despite the missing phonetic transcriptions and while the audio files do not exist for interviews 40–57, there are other levels of representation that do. More crucially, since the audio files for 65–107 survive (though in a degraded state), some enhancement will be possible via digitization and, once this has been done, orthographic transcription can be attempted. We aim, therefore, to incorporate these materials into the corpus once the XSLT phase described immediately above has been completed.

5.2 Preliminary linguistic analyses

One aspect of the NECTE project, which is particularly gratifying, is that it has already begun to further the objectives of the original TLS and PVC research agendas (see section 2.1 and 2.2 above). Preliminary linguistic analyses of various kinds have been performed by the NECTE team on different aspects of the data. We describe below two areas that have proved fruitful in this regard and which we, therefore, intend to pursue further.

5.2.1 Corpus linguistics, dialectology, (historical) (English) linguistics and English language

Previous research prior to the establishment of this resource provided evidence suggesting that certain dialects preserve historical features of English that are no longer extant in standard and other regional varieties. In published research by Beal (2004a, 2004b) and Beal and Corrigan (2002, 2005a, 2005b), arising from the NECTE programme, we have demonstrated that the corpus is an extremely useful tool in this regard. It has permitted us, for example, to track the development of relative clause markers, adverbs and patterns of negation in real time (by comparing speakers across the 1969 and 1994 corpora).

Indeed, we have also begun cross-dialectal investigations (between NECTE and the Corpus of Sheffield Usage (CSU) (Beal, 2002)) that have allowed us to uncover morphosyntactic variation in the system of relativization in non-standard Englishes across regional space (see particularly Beal & Corrigan, 2005b). An important finding of these investigations has been that some language features are similar across the two dialects (such as the use of *nowt* as a negator, discussed in Beal and Corrigan, 2005a), whereas others distinguish dialects quite strikingly

(as does the use of *what* as a relative marker found to be extremely rare in NECTE but fairly common in the CSU (Beal and Corrigan, 2002, 2005b)).

Although we have been fortunate with respect to the congruities between NECTE and the CSU, a real impediment to the advancement of knowledge as regards tracking the development of English more globally has been the lack of standardization with respect to the manner in which electronic vernacular corpora are encoded. As Bauer (2002, pp. 107–8) notes:

> On the whole, corpora have been built for national varieties of English rather than for regional dialects within one country. Thus we do not have public electronic corpora that would allow us to investigate differences in the syntax of Newfoundland and Vancouver Englishes, or of Cornish and Tyneside Dialects.

His motivation for this statement partly arises from the fact that 'diverse collections may be comprised of slightly different types of data'. As such, an important contribution to this subject area that NECTE has made is in the creation of protocols and guidelines regarding the collection, transcription, annotation and long-term preservation of vernacular corpora. When applied by other researchers to their own data sets, these will allow cross-variety comparisons of exactly the sort Bauer is lamenting the lack of. There is evidence of a perceived need internationally for such standards, as argued in Kretzschmar *et al*. (2005, and forthcoming).

5.2.2 Exploratory multivariate analysis

The highly detailed phonetic transcriptions of the Tyneside Linguistic Survey interviews that we have now restored offer a unique opportunity for applying exploratory multivariate analytical techniques, such as cluster and principal components analysis as well as various non-linear methods, to the interrogation of linguistic corpora.

Hence, some members of the NECTE team have begun to develop the empirical methodology based on exploratory multivariate analysis that the TLS used for selection of linguistic variables, described in section 2.1 above. Published work thus far evaluates the reliability of the results generated by hierarchical cluster analysis, the analytical method used by the TLS, and proposes the application of more recently developed methods such as the Self Organizing Map to analyses of the TLS phonetic transcription data (Moisl and Beal 2001; Moisl and Jones,

2005). More recently, in papers presented at ICLAVE 3 (Moisl *et al.*, 2005) and UKLVC 5 (Maguire and Moisl, 2005) and currently being prepared for publication, the TLS results reported in Jones-Sargent (1983) have been replicated and extended. These outputs have demonstrated that the TLS speakers fall into clearly defined groups on the basis of their phonetic usage and that these groups correlated well with the socio-economic backgrounds of individual informants. These results are congruent not only with the interrogation of the corpus at the morphosyntactic level outlined in section 5.2.1 above, but are also interesting from the perspective of the social dynamics of Newcastle speech captured in research by the PVC team some thirty years after the collection of the TLS data.

6 Conclusion

The past two decades or so have resolved many of the corpus creation difficulties that beset the original TLS team in particular, and we have been at pains to enhance both the TLS and PVC corpora in these respects. For data entry, verification and correction we have made use of optical character recognition, graphical user interfaces, text processing systems, and a variety of text analysis and diagnostic software. For standards, we have adopted XML and TEI. Our plans for dissemination have moved considerably beyond what might have been imagined by either of the original research teams whose data we inherited, in that we have taken full advantage of the connectivity of the internet and the ever-developing facilities of the Web. It is now, in fact, possible to construct electronic corpora in the manner in which the TLS project intended, and to publish it as a resource for the research community in a way that its members did not, and could not, have conceived.

Appendix

Table 2.A1 Existing TLS source materials

Interview number	Tape exists	Index card set exists	Electronic phonetic transcription file exists	Social data file exists
1	X	X	X	X
2	X	X	X	X
3	X	X	X	X
4	X	X	X	X
5	X	X	X	X
6	X	X	X	X

Table 2.A1 Continued

Interview number	Tape exists	Index card set exists	Electronic phonetic transcription file exists	Social data file exists
7	X	X	X	X
8	X	X	X	X
9	X	X	X	X
10	X	X	X	X
11	X	X	X	X
12	X	X	X	X
13	X	X	X	X
14	X	X	X	X
15	X	X	X	X
16	X	X	X	X
17	X	X	X	X
18	X	X	X	X
19	X	X	X	X
20	X	X	X	X
21	X	X	X	X
22	X	X	X	X
23	X	X	X	X
24	X	X	X	X
25	X	X	X	X
26	X	X	X	X
27	X	X	X	X
28	X	X	X	X
29	X	X	X	X
30	X	X	X	X
31	X	X	X	X
32	X	X	X	X
33	X	X	X	X
34	X	X	X	X
35	X	X	X	X
36	X	X	X	X
37	X	X	X	X
38	X	X		X
39	X	X		X
40		X	X	X
41		X	X	X
42		X	X	X
43		X	X	X
44		X	X	X
45		X	X	X
46		X	X	X
47		X	X	X
48		X	X	X
49		X	X	X
50		X	X	X

Table 2.A1 Continued

Interview number	Tape exists	Index card set exists	Electronic phonetic transcription file exists	Social data file exists
51		X	X	X
52		X	X	X
53		X	X	X
54		X	X	X
55		X	X	X
56		X	X	X
57		X	X	X
58			X	X
59			X	X
60			X	X
61			X	X
62			X	X
63			X	X
64			X	X
65	X			
66	X			
67	X			
68	X			
69	X			
70	X			
71	X			
72	X			
73	X			
74	X			
75	X			
76	X			
77	X			
78	X			
79	X			
80	X			
81	X			
82	X			
83	X			
84	X			
85	X			
86	X			
87	X			
88	X			
89	X			
90	X			
91	X			
92	X			
93	X			
94	X			

Table 2.A1 Continued

Interview number	Tape exists	Index card set exists	Electronic phonetic transcription file exists	Social data file exists
95	X			
96	X			
97	X			
98	X			
99	X			
100	X			
101	X			
102	X			
103	X			
104	X			
105	X			
106	X			
107	X			
108	blank			
109	blank			
110	damaged			
111	damaged			
112	blank			
113	damaged			
114	blank			

Notes

1. The authors would like to acknowledge the financial support of the Arts and Humanities Research Board (AHRB) (grant no: RE11776) in funding this resource enhancement project entitled: A Linguistic 'Time-Capsule': The Newcastle Electronic Corpus of Tyneside English. We also appreciate the helpful comments generated by an oral version of this chapter delivered at the Models and Methods panel, which took place at Sociolinguistics Symposium 15, University of Newcastle, April 2004.
2. As well as preserving the data from the TLS project, the NECTE team has revisited its methodology and sought to devise new methods capable of achieving the original aims of the TLS team (Allen *et al*., 2003a, 2003b; and Jones and Moisl, 2003, for instance). As our preliminary analyses in section 5 demonstrate, the much greater computer power available today has allowed us to implement computationally more demanding cluster analysis algorithms, such as self-organizing maps, which could not have run within a reasonable time span in the 1960s, and they have already produced quite remarkable results.
3. In this regard, there are some significant publications arising from the research, including: Local (1982), Jones (1985), and Local *et al*. (1986), as well as those mentioned in section 1 above.

4. Requests for access since the release of the corpus in July 2005 suggest that NECTE is also valuable to linguists in the fields of discourse analysis and language and gender, given the informal and unstructured nature of the dyadic interaction captured in the data and the mixture of same and differently gendered pairs of subjects. Scholars from other disciplines (folklore and history, for example) have also shown interest and it is expected that the impact of the corpus will increasingly be more wide-ranging as it becomes better-known outside English language and linguistics.
5. These materials are more fully described than we have space for in Jones-Sargent (1983).
6. We are grateful to Jonathan Marshall of the University of Gloucester for his assistance with the acoustic filtering procedures.
7. The NECTE amalgamation scheme also included these, partly for preservation purposes. Electronic copies of the orthographic transcription text on the index cards were made, and the copies were proofread relative to the cards. No changes of any kind, including corrections, were made as per normal practice in linguistic archaeology (see Meurman-Solin, this volume). The reader should also note that the TLS team only ever transcribed the interviewees' utterances, ignoring the interviewer entirely, though this has not been NECTE's practice.
8. Because it was specifically designed for handling Standard English text, there is no guarantee that tagging accuracy comparable to that for the BNC has been achieved for NECTE using the CLAWS software. We have, however, performed an amount of subsequent proofreading and found the error rate to be not unduly high. Specific accuracy levels, of course, remain to be determined by subsequent detailed study of this level of the corpus, which is beyond the scope of the NECTE project. For further details on the software itself than we have space for here, see http://www.comp.lancs.ac.uk/computing/research/ucrel/
9. The reader should be aware that the specifics of which numbers are used in the code are irrelevant in each case, and could have been anything else.
10. It is crucial to note that the Gateshead TLS transcriptions were done exclusively by a single member of the project, Vince McNeany, who was both a trained phonetician and a native speaker of the Tyneside dialect. This is important for subsequent analyses of the phonetic level because it minimizes the subjectivity and variation that inevitably compromises phonetic transcriptions. It still remains unclear who exactly undertook the Newcastle transcriptions and this may significantly impact upon their reliability, in comparison with the Gateshead sample.
11. We should point out that no attempt has been made by the NECTE team either to review the TLS phonetic transcriptions relative to the original audio recordings, or to extend/further refine the phonetic representation to accommodate what the TLS did not originally encode. The TLS transcriptions are, rather, offered as an historical artefact, and the reason they are included in NECTE is principally because of their intrinsic interest to researchers who want to study the phonetics of the TLS material. The phonetic analysis encoded is extremely detailed (much more so than that of current practice within auditory sociophonetic research, for instance), providing from one to ten realizations of any given phonological

segment, and this will no doubt be extremely useful to certain kinds of end-user.

12. It is, of course, a straightforward matter to decrease granularity by multiples of 20 if required. Finer granularity would, however, require insertion of markers at the appropriate places in all levels of representation and while this is possible, of course, it requires considerable additional human intervention.
13. An online version can be viewed at: http://www.tei-c.org
14. Further details of the TEI-conformant XML encoding are available from the NECTE website: http://www.ncl.ac.uk/NECTE/index.htm
15. Some additional directories of XML-aware software include: _http://xml.coverpages.org/publicSW.html; http://www.xmlsoftware.com/__http://www.garshol.priv.no/download/xmltools/__http://www.wdvl.com/Software/XML/

References

Allen, W., J. C. Beal, K. P. Corrigan, H. Moisl and C. Rowe. 2003a. 'A linguistic "time-capsule": *The Newcastle Electronic Corpus of Tyneside English*'. Poster presented at the 2nd International Conference on Language Variation and Change in Europe, University of Uppsala, June 2003.

Allen, W., J. C. Beal, K. P. Corrigan, H. Moisl and C. Rowe. 2003b. 'A linguistic "time-capsule": *The Newcastle Electronic Corpus of Tyneside English*'. Website display presented at NWAVE, University of Philadelphia, October 2003.

Bauer, L. 2002. 'Inferring variation and change from public corpora'. *The Handbook of Language Variation and Change*, ed. by J. K. Chambers, P. Trudgill and N. Schilling-Estes, pp. 97–114. Oxford: Blackwell.

Beal, J. C. 1994–95. *The Catherine Cookson Archive of Northumbrian Dialect*. Catherine Cookson Foundation.

Beal, J. C. 2002. *The Corpus of Sheffield Usage*. British Academy Small Research Grant.

Beal, J. C. 2004a. 'The phonology of English dialects in the North of England'. *A Handbook of Varieties of English*, Volume I, ed. by B. Kortmann, pp. 113–33. Berlin: Mouton.

Beal, J. C. 2004b. 'The morphology and syntax of English dialects in the North of England'. *A Handbook of Varieties of English*, Volume II, ed. by B. Kortmann, pp. 114–41. Berlin: Mouton.

Beal, J. C. 2004c. 'Geordie nation: language and identity in the North-East of England'. *Lore and Language* 17:33–48.

Beal, J. C. 2005. 'Dialect representation in texts'. *The Encyclopedia of Language and Linguistics*, 2nd edn, ed. by K. R. Brown, pp. 351–8. Oxford: Elsevier.

Beal, J. C. and K. P. Corrigan. 1999. *Investigating the Social Trajectories of Modal Verb Usage in Tyneside English*. Newcastle University Research Committee, Vacation Scholarship Panel.

Beal, J. C. and K. P. Corrigan. 2000a. 'A dynamic re-modelling of linguistic variation: the social trajectories of syntactic change amongst young Tynesiders,

1969–1994'. Paper presented at the Sociolinguistics Symposium, University of the West of England, Bristol, April 2000.

Beal, J. C. and K. P. Corrigan. 2000b. 'The Newcastle–Poitiers Electronic Corpus of Tyneside English'. Paper presented at the 11th International Conference on English Historical Linguistics, University of Santiago de Compostella, August 2000.

Beal, J. C. and K. P. Corrigan. 2000c. 'New ways of capturing the "Kodak moment": Real-time vs. apparent time analyses of syntactic variation in Tyneside English, 1969–1994'. Paper presented at the 2nd Variation is Everywhere Workshop, University of Essex, September 2000.

Beal, J. C. and K. P. Corrigan. 2000–01. *The Newcastle–Poitiers Corpus of Tyneside English*. British Academy Small Grant no.: SG-30122.

Beal, J. C. and K. P. Corrigan. 2002. 'Relativisation in Tyneside and Northumbria'. *Relativisation on the North Sea Litoral* (LINCOM Studies in Language Typology, 7), ed. by P. Poussa, pp. 125–34. München: Lincom Europa.

Beal, J. C. and K. P. Corrigan. 2005a. '"No, nay, never", negation in Tyneside English'. *Aspects of English Negation*, ed. by Y. Iyeiri, pp. 139–56. Tokyo: Yushodo University Press, and Amsterdam: John Benjamins.

Beal, J. C. and K. P. Corrigan. 2005b. 'A tale of two dialects: relativisation in Newcastle and Sheffield'. *Dialects Across Borders: Selected Papers from the 11th International Conference on Methods in Dialectology (Methods XI), Joensuu, August 2002*, *CILT, 273*, ed. by M. Filppula, J. Klemola, M. Palander and E. Penttilä, pp. 211–29. Amsterdam: John Benjamins.

Beal, J. C., K. P. Corrigan, J.-L. Duchat, M. Fryd and C. Gérard. 1999–2000. *The Newcastle–Poitiers Corpus of Tyneside English*. British Academy and British–French Joint Projects with the Centre National de la Recherche Scientifique (CNRS), RSU Code: RES/3300/7001.

Brockett, J. T. 1825. *A Glossary of North Country Words, in Use*. Newcastle upon Tyne: E. Charnley.

Cameron, D. 2001. *Working with Spoken Discourse*. London: Sage.

Corrigan, K. P. 1999-2000. *Syntactic Change in Progress? The Newcastle–Poitiers Electronic Corpus of Tyneside English*. Newcastle University Research Committee, Vacation Scholarship Panel.

Corrigan, K. P., H. Moisl and J. C. Beal. 2001–05. *A Linguistic 'Time-Capsule': The Newcastle Electronic Corpus of Tyneside English*. Arts and Humanities Research Board (AHRB), Grant no.: RE11776 (http://www.ncl.ac.uk/NECTE).

Dobson, S. 1974. *The New Geordie Dictionary*. Newcastle: Frank Graham.

Docherty, G. and P. Foulkes. 1999. 'Derby and Newcastle: instrumental phonetics and variationist studies'. *Urban Voices: Accent Studies in the British Isles*, ed. by P. Foulkes and G. Docherty, pp. 47–71. London: Arnold.

Douglas, P. 2001. *Geordie–English Glossary*. London: Abson Books.

Geeson, C. 1969. *A Northumberland and Durham Word Book: The Living Dialect, Including a Glossary, with Etymologies and Illustrative Quotations, of Living Dialect Words*. Newcastle upon Tyne: H. Hill.

Graham, F. (ed.). 1979. *The New Geordie Dictionary*. Newcastle: Frank Graham.

Griffiths, B. 1999. *North-East Dialect: Survey and Word-List*. Gateshead: Athenaeum Press.

Heslop, R. O. 1892–94. *Northumberland Words: A Glossary of Words Used in the County of Northumberland and on the Tyneside*. London: English Dialect Society.

Jones, V. 1985. 'Tyneside syntax: a presentation of some data from the Tyneside Linguistic Survey'. *Focus on England and Wales*, ed. by W. Viereck, pp. 163–77. Amsterdam: John Benjamins.

Jones, V. and H. Moisl. 2003. 'Cluster analysis of the *Newcastle Electronic Corpus of Tyneside English*: a comparison of methods'. Paper presented at Web X: A Decade of the World Wide Web, Joint International Conference for the Association for Literary and Linguistic Computing, University of Georgia, Athens, Georgia, May–June 2003.

Jones-Sargent, V. 1983. *Tyne Bytes: A Computerised Sociolinguistic Study of Tyneside*. Frankfurt am Main: Peter Lang.

Kerswill, P. and S. Wright. 1990. 'The validity of phonetic transcription: limitations of a sociolinguistic research tool'. *Language Variation and Change* 2:225–75.

Kirk, J. 1997. 'Irish-English and contemporary literary writing'. *Focus on Ireland*, ed. by J. Kallen, pp. 190–205. Amsterdam: John Benjamins.

Kretzschmar, W. A. Jr, J. Anderson, J. C. Beal, K. P. Corrigan, L. Opas-Hänninen and B. Plichta. 2005. 'Collaboration on corpora for regional and social analysis'. Paper presented at AACL 6/ICAME 26, University of Michigan, Ann Arbor, May 2005.

Kretzschmar, W. A. Jr, J. Anderson, J. C. Beal, K. P. Corrigan, L. Opas-Hänninen & B. Plichta. (forthcoming). 'Collaboration on corpora for regional and social analysis'. Special Issue of *Journal of English Linguistics*.

Labov, W. 1972. *Sociolinguistic Patterns*. Philadelphia: Pennsylvania University Press.

Lawrence, H., S. A. Tagliamonte and J. Smith. 2003. 'Transcription technicalities'. Paper presented to the NECTE workshop 'Deriving Standards for the Creation of Electronic Vernacular Corpora: Tagging and Transcription Issues' at the Fourth UK, Language Variation and Change conference, University of Sheffield, September 2003.

Local, J. K. 1982. 'How many vowels in a vowel?' *Journal of Child Language* 10:449–53.

Local, J. K., J. Kelly and W. H. G. Wells. 1986. 'Towards a phonology of conversation turn-taking in Tyneside'. *Journal of Linguistics* 22:411–37.

Macaulay, R. K. S. 1991. '"Coz it izny spelt when they say it": displaying dialect in writing'. *American Speech* 66:280–91.

Macaulay, R. K. S. 2005. *Talk That Counts: Age, Gender and Social Class Differences in Discourse*. Oxford: Oxford University Press.

Maguire, W. and H. Moisl. 2005. 'Identifying the main determinants of phonetic variation in the Newcastle Electronic Corpus of Tyneside English'. Paper presented to the Fifth UK Language Variation and Change conference, Aberdeen, September 2005.

Milroy, L. 1984. 'Urban dialects in the British Isles'. *Language in the British Isles*, ed. by P. Trudgill, pp. 199–218. Cambridge: Cambridge University Press.

Milroy, J., L. Milroy and G. Docherty. 1997. 'Phonological variation and change in contemporary spoken British English'. ESRC, Unpublished Final Report, Dept. of Speech, University of Newcastle upon Tyne.

Moisl, H. and J. C. Beal. 2001. 'Corpus analysis and results visualization using self-organizing maps'. *Proceedings of the Corpus Linguistics 2001 Conference, UCREL Technical Papers 13 – Special Issue*, ed. by P. Rayson,

A. Wilson, T. McEnery, A. Hardie and S. Khoja, pp. 386–91. UCREL: Lancaster University.

Moisl, H. and V. Jones. 2005. 'Cluster analysis of the Newcastle Electronic Corpus of Tyneside English: a comparison of methods'. *Literary and Linguistic Computing* 20:125–46.

Moisl, H., W. Maguire and W. Allen. 2005. 'Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English'. Paper presented to the Third International Conference on Language Variation in Europe, Amsterdam, June 2005.

Moody, T. (forthcoming) *Glossary of Tyneside and Northumbrian English*, ed. by J. C. Beal and K. P. Corrigan. Newcastle: Northumbrian Language Society.

Pellowe, J. and V. Jones. 1978. 'On intonational variety in Tyneside speech'. *Sociolinguistic Patterns in British English*, ed. by P. Trudgill, pp. 101–21. London: Arnold.

Pellowe, J., B. H. M. Strang, G. Nixon and V. McNeany. 1972. 'A dynamic modelling of linguistic variation: the urban (Tyneside) linguistic survey'. *Lingua* 30:1–30.

Poplack, S. 1989. 'The care and handling of a megacorpus: the Ottowa–Hull French Project'. *Language Change and Variation,* ed. by R. Fasold & D. Schiffren, pp. 411–51. Amsterdam: Benjamins.

Preston, D. 1985. 'The Li'l Abner syndrome: written representations of speech'. *American Speech* 60(4):328–36.

Sacks, H., E. Schegloff and G. Jefferson. 1974. 'A simplest systematics for the organization of turn-taking for conversation'. *Language* 50(4):696–735.

Sperberg-McQueen, C. and L. Burnard (eds). 2002. *Guidelines for Text Encoding and Interchange*. Published for the TEI Consortium by the Humanities Computing Unit, University of Oxford (http://www.tei-c.org)

Stenstrom, A.-B. and J. Svartvik. 1994. 'Imparsable speech: repeats and nonfluencies in spoken English'. *Corpus-based Research into Language*, ed. by N. Oostdijk and P. de Haan, pp. 241–54. Amsterdam: Rodopi.

Strang, B. 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. 2006. 'Representing real language: consistency, trade-offs and thinking ahead!' *Creating and Digitizing Language Corpora: Synchronic Databases (Volume 1)*, ed. by Joan C. Beal, Karen P. Corrigan and Hermann Moisl, pp. 000. Basingstoke: Palgrave Macmillan.

Trudgill, P. 1974. *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.

van den Broek, T. 2004. *Benchmarking XML-editors*. Electronic Publication, Arts and Humanities Data Service. http://ahds.ac.uk/creating/information-papers/xml-editors/

Watt, D. 2002. '"I don't speak with a Geordie accent, I speak, like, the Northern accent": contact induced levelling in the Tyneside vowel system'. *Journal of Sociolinguistics* 6(1):44–63.

Watt, D. and L. Milroy. 1999. 'Patterns of variation in Newcastle vowels', *Urban Voices: Accent Studies in the British Isles*, ed. by P. Foulkes and G. Docherty, pp. 25–46. London: Arnold.

Wells, J. 1982. *Accents of English I. An Introduction*. Cambridge: Cambridge University Press.

Widdowson, J. D. A. 2003. 'Hidden depths: exploiting archival resources of spoken English'. *Lore and Language* 17(1/2):81–92.

Websites

CLAWS4, part-of-speech tagger for English (UCREL): http://www.comp.lancs.ac.uk/computing/research/ucrel/
NECTE: http://www.ncl.ac.uk/NECTE
oXygen XML editor: http://www.oxygenxml.com/
TEI Guidelines: http://www.tei-c.org (see Sperberg-McQueen & Burnard, 2002)
Xaira (Oxford University Computing Service): http://www.oucs.ox.ac.uk/rts/xaira/
XML (Extensible Markup Language): http://www.w3.org/XML/
XSLT (Extensible Stylesheet Language Transformations): http://www.w3.org/TR/xslt