The Newcastle Electronic Corpus of Tyneside English: Annotation Practices and Dissemination Strategies.[1]

Joan C. Beal, University of Sheffield.

Karen P. Corrigan, Newcastle University.

Hermann L. Moisl, Newcastle University.

Abstract

This chapter describes the construction of the Newcastle Electronic Corpus of Tyneside English (NECTE), a legacy corpus based on data collected for two sociolinguistic surveys conducted on Tyneside in the north-east of England in c.1969 and 1994, respectively. It focusses on transcription issues relevant for addressing research questions in phonetics/phonology. There is also discussion of the rationale for the text encoding systems adopted in the corpus construction phase as well as the dissemination strategy employed since completion in 2005.

Key Words

Tyneside Coding Transcription Tagging Dissemination

1.0 Introduction

The *Newcastle Electronic Corpus of Tyneside English* (NECTE) is a corpus based on legacy materials originally collected for sociolinguistic surveys in Tyneside, north-east England (Kretzschmar et al. 2006). It is unique from the perspective of other large-scale, web-based corpora in a number of respects. Firstly, although it is true that the

[1] The authors would like to acknowledge the financial support of the Associated Humanities Research Board (AHRB) (grant no: RE11776) in funding the resource enhancement project entitled: *A Linguistic 'Time-Capsule': The Newcastle Electronic Corpus of Tyneside English* described in this chapter.

original intention of the data collection initiatives which form its bedrock did not envisage that the data would ever become a publicly-available electronic corpus, the aim was wholly phonetic/phonological as demonstrated by the outputs which resulted from the original research (see Docherty & Foulkes, 1999; Jones-Sargent, 1983; Local, 1982; Local et al. 1986; Milroy et al. 1994/1997; Pellowe et al. 1972; Pellowe & Jones, 1978; Strang, 1968; Watt 2002 and Watt & Milroy 1999). Secondly, NECTE was designed for wide-ranging free distribution and incorporates numerous levels of annotation (including phonetic/phonological) which adhere to world standards. This chapter will describe the construction of the NECTE corpus, focussing on transcription issues relevant for addressing research questions in phonetics/phonology. There will also be discussion of the rationale for the text encoding systems adopted in the corpus construction phase as well as the dissemination strategy employed since completion in 2005. The exploitation of NECTE for phonetic / phonological analysis is described by Moisl's chapter in Part I of this Handbook.

2.0 NECTE Corpus Construction

The NECTE project aimed to improve access to and promote the re-use of Tyneside recordings from the twentieth century. The objective was to create the first electronic vernacular corpus that was aligned, part-of-speech tagged and fully compliant with international standards for encoding text. The resultant public corpus amalgamated two collections, the *Tyneside Linguistic Survey* (TLS) (Strang 1968) and the *Phonological Variation and Change in Contemporary Spoken English* (PVC) project (Milroy et al. 1997), into a single dataset formatted in Text Encoding Initiative (TEI)-conformant XML (see http://www.tei-c.org/index.xml). It is documented in and downloadable from the project website (http://www.ncl.ac.uk/necte/) and is compatible with XML-aware

analytical software. It was intended primarily as a resource for linguistic/(socio-)historical researchers and has since been exploited to address a range of questions in research of different kinds with a phonetics/phonological orientation (see, for instance, Maguire 2007, Moisl et al. 2006 and Moisl & Maguire 2008).[2]

2.1 NECTE Orthographic and Phonetic Transcription Practices

Given exigencies of space, the account offered here with respect to the principles and methods underpinning the transcription practices to which NECTE conforms is necessarily brief. More detailed general treatments can be found in Allen et al. (2007) as well as Beal, Corrigan, Smith & Rayson (2007). The specific issues relating to the phonetic transcriptions associated with the TLS sub-corpus are outlined in Moisl's contribution in Part I of this volume.

As noted above, the audio content of the TLS and PVC corpora which was retrievable[3] is included in NECTE and these spoken materials were transcribed using orthographic representations that matched as far as possible the conventions of Standard British English and used a strict orthographic transcription protocol (OTP) to ensure consistency. This decision arose from consideration of research discrediting the practice of representing non-standard phonology using semi-phonetic spelling (Preston 1985, 2000) and other more recent work on good practice for "representing real language" (Tagliamonte 2005). As the latter notes, this involves 'trade-offs' of various kinds. For instance, standard conventions were not followed if the item was lexically or morphologically distinct. Hence, while the characteristic Tyneside pronunciation of /na:/

[2] It has also been used for other types of linguistic analyses and, indeed, for research in fields as diverse as European ethnology, world heritage studies, multivariate analysis and web design (see Allen et al. 2005; Anderwald & Szmrecsanyi 2009; Beal 2009; Beal & Corrigan 2007; Beal & Corrigan 2009 and Moisl & Jones 2005 for a selection).

[3] Given the fact that the TLS dates from the 1960's, coupled with the fact that the material was not systematically catalogued prior to NECTE, some of the original projected recordings associated with this initiative were either missing or had deteriorated to the point were digitization was no longer possible (see Allen et al. 2007 and Beal 2009).

for SE *know* is often spelt <knaa> in popular representations of the dialect, NECTE transcribes the lexeme as <know>. By contrast, the typical negative form of *do* amongst certain groups of speakers on Tyneside (particularly in vernacular style) was represented as <divn't> rather than the more conventional <don't> since it is both lexically and morphologically distinctive and it was felt important not to ignore this contrast.

Naturally, as a corpus of dialectal English, the dataset contains numerous lexemes for which there is no standard equivalent. Where these already had an agreed spelling in, for example, Heslop (1893-1894), that convention was adopted in NECTE's OTP. Hence, *gan* for 'go' is in Helsop (1893-1894: 315) and was used with this spelling in NECTE. Other lexemes in the TLS/PVC interviews which were not found in such reliable dialect dictionaries (often because the latter are historical rather than contemporary) had forms with semi-phonetic spellings adapted for them (like <happy baccy> for 'cannabis' and others cited in Appendix 2 of NECTE available to download from http://www.ncl.ac.uk/necte/appendix2.htm).

There were 'trade-offs' too in the phonetic transcriptions accompanying NECTE from those originally envisaged when the project was first conceived. In the first place, consultation with both the original PVC project team and with other (socio-)phoneticians suggested that while it seemed sensible to provide samples of phonetically transcribed data from the PVC audio files, we were encouraged not to provide full sets of transcripts for all the interviews on the basis that end users with a phonetics/phonological orientation will generally prefer to undertake their own transcriptions (see Kerswill & Wright 1990). Moreover, the advent of ELAN and PRAAT and other software tools such as those detailed in Part III of this volume, mean that more sophisticated analysis than the rather rudimentary auditory transcriptions of the samples we have provided are eminently possible with this sub-corpus of NECTE. The situation proved rather trickier

with the older TLS recordings as a direct consequence of the fact (noted elsewhere) that these are legacy materials of some considerable vintage. Crucially, from a phonetics/phonology perspective, the original audio recordings were already accompanied by a detailed phonetic transcription and rudimentary part-of-speech tagging as demonstrated in the handwritten sample index card in Figure 1:

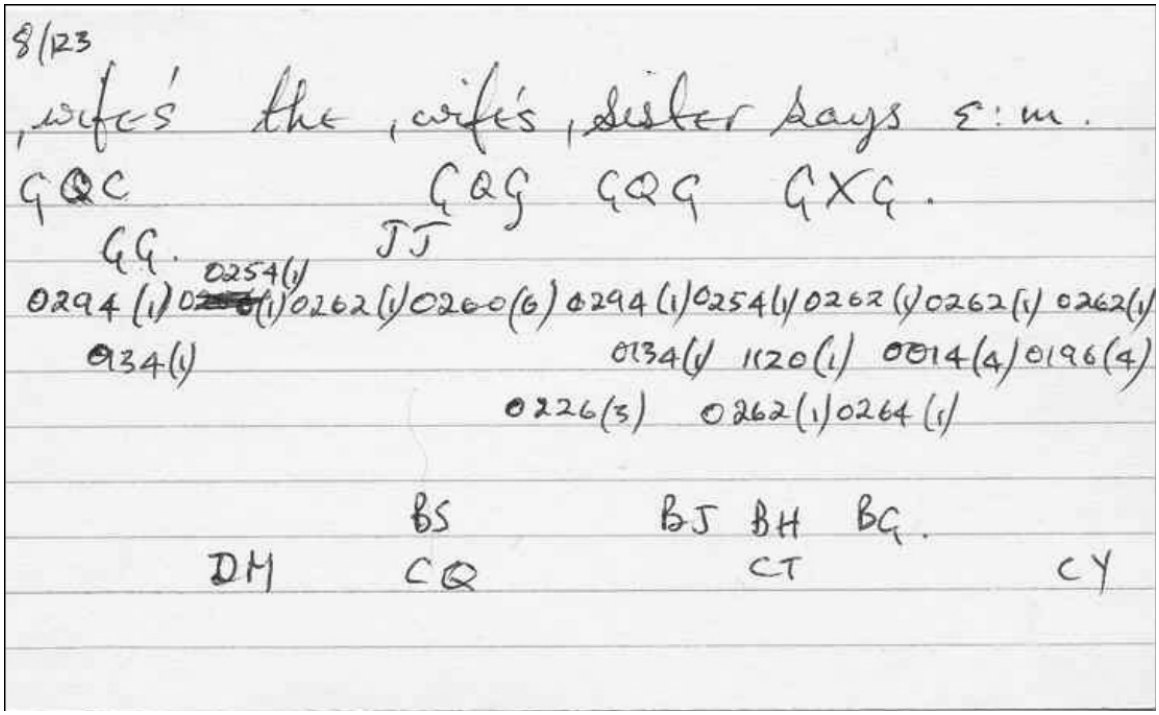

Figure 1: TLS Transcription Card with annotations

It was our intention, therefore, to preserve these transcriptions and to considerably augment the morphosyntactic detail by providing state-of-the-art grammatical tagging for both the TLS and PVC sub-corpora. As regards the former, as can be seen from the numbers on the card in Figure 1, which are, in fact, the phonetic transcriptions, the TLS team eschewed the more conventional IPA system for transcribing and, instead, created a meticulous hierarchical coding system downloadable from

http://www.ncl.ac.uk/necte/appendix1.htm, an extract of which is illustrated in Figure 2.

Figure 2:

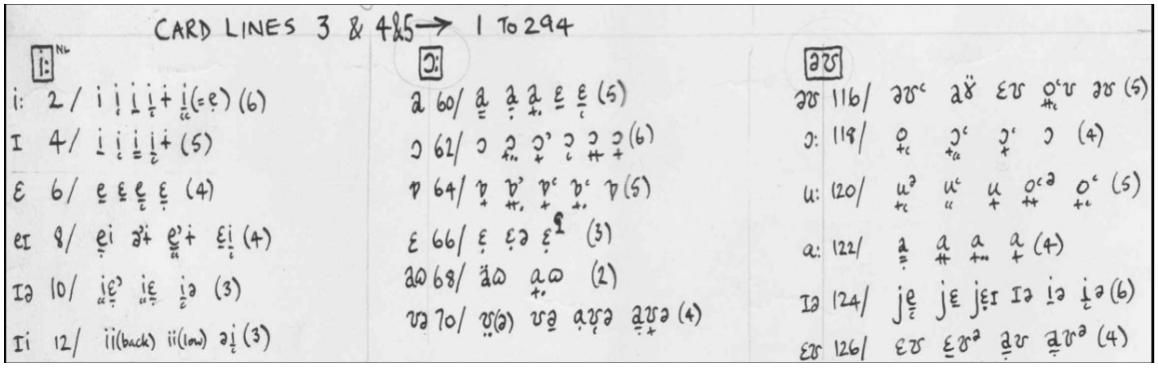

Figure 2: Extract form TLS Coding Scheme

This involves three levels of phonetic/phonological analysis. The boxes at the top of this image relate to Overall Units (OUs), equivalent to the lexical sets used by Wells (1982a/b) so as to facilitate the comparison of English accents globally. The next level is that of the Putative Diasystemic Variants (PDV) which are represented by the IPA symbols in the left-hand column under each OU, and are roughly equivalent to the phonemic level of transcription. The symbols which appear to the right of each PDV are 'states', each of which represents a different phonetic variant and have unique identifier numbers of the kind already illustrated in Figure 1, such that the code for any output indicates not only its precise phonetic nature but also the phoneme of which it is an allophone and the lexical set in which it was used. Given the interest within the TLS team in harnessing computational tools for analyzing variation and change in the phonology of Tyneside English (described more fully in Moisl's contribution to Part I), the transcriptions in Figure 1 were transferred to electronic format.

Initially, from NECTE's perspective, the resultant files, which had been lodged with

the Oxford text Archive in the distant past and were retrieved at the outset of NECTE appeared to be a labour- and time-saving alternative to keying in the numerical codes from the index cards so as to reinstate the phonetic transcriptions which could then be aligned with the sound files, orthographic transcriptions and part-of-speech tagging in the enhanced NECTE corpus. However, a peculiarity that stems from the original electronic data entry system used by the computing staff who had input the data from the TLS team's original index cards meant that the resulting files had to be extensively edited. The problem arose from the way in which the five-digit codes were laid out by the TLS researchers on the index cards as illustrated in Figure 1. For reasons that are no longer clear, all the consonant codes (beginning (0294(1)) in line 4) were written on one line, and all of the vowel codes appear on the line below ((0134(1)) on line 5). When the TLS gave these 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 those for vowels. 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 that phonetic environment had already been captured by their hierarchical scheme), it was highly problematic for the NECTE enhancement of the original materials. Maintaining 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 electronic files were therefore edited with reference to the original index cards so as to restore the correct code sequencing, and the result was proofread for accuracy. The example in Figure 3 shows the intermediate (PDV) TLS phonetic representation – equivalent to a broad segmental phonetic IPA representation. In the corpus, each PDV segment is, however, indexed into up to 10 state variants – equivalent to a (very) detailed phonetic IPA representation, the like of which would hardly ever be contemplated today.

Orthographic	Down by Clark Chapman's
Segmental Phonetic (PDV)	dɛʊn baɪ klɑk ʧæpmənz

Figure 3: Example of NECTE transcriptions

2.2 Part-of-Speech Tagging

As noted elsewhere, grammatical tagging was crucial to the *NECTE* research 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 proof-reading.

The CLAWS tagger, developed for annotating the BNC by the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University, UK, was selected because it fulfilled *NECTE*'s requirements as 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. [4] The *NECTE* orthographic transcriptions of the *TLS* and the

[4] 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

PVC audio were part-of-speech tagged by the CLAWS4 tagger using the UCREL C8 tagset and Figure 4 below 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.*

CLAWS software. We have, however, performed an amount of subsequent proof-reading and found the error rate to be not unduly high. For further details on the software itself than we have space for here, see: Beal et al. (2007); http://www.ncl.ac.uk/necte/appendices.htm - appendix 3 and http://www.comp.lancs.ac.uk/computing/research/ucrel/.

Figure 4: *CLAWS output*

```
<u who="informantTLSg37">
<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>
```

2.3 Coding

In order to ensure NECTE's sustainability in the longer term and to preserve the important metadata we had uncovered during the enhancement of both the TLS and PVC, it was encoded using Text Encoding Initiative (TEI)-conformant Extended Markup Language (XML) syntax. XML (http://www.w3.org/XML/) encourages the creation of information resources that are independent both of the specific characteristics of the computer platforms on which they reside and of the software applications used to interpret them by providing a range of standard document structuring mechanisms which are adaptable to particular applications. TEI proposes a selection of XML constructs specific to the creation of natural language corpora, and thereby defines a standard for corpus construction that has been widely adopted. This section outlines the TEI-conformant XML structure of NECTE.

Every TEI corpus consists of two main elements: (i) a prolog that contains meta-information about the corpus, and (ii) the document instance that contains the content of the corpus. This is shown in figure 4.

```
<teiCorpus.2>
        <teiHeader type='corpus'>
                <fileDesc></fileDesc>
                <encodingDesc></encodingDesc>
                <profileDesc></profileDesc>
                <revisionDesc></revisionDesc>
        </teiheader>
        &tlsg01;&tlsg22;&tlsn06; &tlsg02;&tlsg23;&tlsn07; &tlsg03;&tlsg24;&pvc01; &tlsg
        04;&tlsg25;&pvc02; &tlsg05;&tlsg26;&pvc03; &tlsg06;&tlsg27;&pvc04; &tlsg07;&tl
        sg28;&pvc05; &tlsg08;&tlsg29;&pvc06; &tlsg09;&tlsg30;&pvc07; &tlsg10;&tlsg31;
        &pvc08; &tlsg11;&tlsg32;&pvc09; &tlsg12;&tlsg33;&pvc10; &tlsg13;&tlsg34;&pvc
        11; &tlsg14;&tlsg35;&pvc12; &tlsg15;&tlsg36;&pvc13; &tlsg16;&tlsg37;&pvc14; &t
```

```
        lsg17;&tlsn01;&pvc15; &tlsg18;&tlsn02;&pvc16; &tlsg19;&tlsn03;&pvc17; &tlsg20;
        &tlsn04;&pvc18; &tlsg21;&tlsn05;
</ teiCorpus.2>
```

Figure 4: TEI structure of NECTE

XML is based on tags enclosed in angle-brackets, and the fundamental principle is that a tag, once opened by <xxx>, must be closed by a corresponding closing bracket </xxx>; tag-pairs can contain other tag-pairs to any depth of embedding, with the result that a TEI corpus has the structure of a context-free phrase structure tree. In Figure 4 the root of the tree is the *<teiCorpus.2> / </ teiCorpus.2>* tag pair representing the entire corpus, and it contains a *<teiHeader type='corpus'> / </teiheader>* subtree representing the prolog (which itself contains subtrees) and a collection of '&******' strings, each of which is, in TEI-speak, an 'entity reference' that refers to a file containing a single NECTE speaker interview.

The *<teiHeader type='corpus'> / </teiheader* prolog in Figure 4 contains a range of metadata relating to NECTE which is subcategorized into several subtrees represented by the constituent tag pairs. This metadata is too complex to be described here, though it is available at the NECTE website. Hence, the remainder of this section outlines the structure of the entity references.

Each entity reference like "&tlsg01" contains a single interview whose structure is shown in Figure 5.

```
<TEI.2 id="tlsg01">
  <teiHeader type="text">
      <!--Header information -->
  </teiHeader>
  <text>
```

```
        <!-- Content -->
    </text>
</TEI.2>
```

Figure 5: The structure of a NECTE speaker interview

The <*TEI.2 id="tlsg01"*> / </*TEI.2*> tree represents the interview named as "tlsg01" in the opening tag, and the <*teiHeader type="text"*> / </*teiHeader*> subtree contains metadata specific to that interview; the <*text*> / </*text*> subtree represents the interview content, the structure of which is given in Figure 6.

```
<text>
    <group>
        <text id='tlsg01audio'>
            <body>
                <!-- content -->
            </body>
        </text>
        <text id='tlsg01necteortho'>
            <body>
                <!-- content -->
            </body>
        </text>
        <text id='tlsg01phonetic'>
            <body>
                <!-- content -->
            </body>
        </text>
        <text id='tlsg01tagged'>
            <body>
                <!-- content -->
            </body>
        </text>
    </group>
</text>
```

Figure 6: The structure of the NECTE interview content

The <*text*> / </*text*> subtree groups several types of content, described earlier, using the

<group> / *</group>* tag pair, and each type of content is represented by a *<text>* / *</text>* subtree: *<text id='tlsg01audio'>* contains audio material, *<text id='tlsg01necteortho'>* contains the orthographic transcription of the audio, *<text id='tlsg01phonetic'>* the phonetic transcription of the audio, and so on. Figure 7 gives an impression of what the structure in Figure 6 looks like in practice.

```
<text>
    <group>
        <text id="tlsg01audio">
            <body>
                <p>tlsg01 audio file</p>
                    <audio entity="tlsaudiog01" />
            </body>
        </text>
        <text id="tlsg01necteortho">
            <body>
                <u who="interviewerTlsg01">
                    <anchor id="tlsg01necteortho0000" />
                    ehm well could you tell us first of all where you were born please where you born in gateshead...
                </u>
                ….Remainder of orthographic representation…
            </body>
        </text>
        <text id='tlsg01phonetic'>
            <body>
                <u who="interviewerTlsg01">
                    <anchor id="tlsg01necteortho0000" />
                    02441 01123 02301 02621 02363 02741 02881 00906 02081 02301
                    02322 01443 02741 02201 01284 02383 02801 00421 02421 02501
                    00342 02164 02721 02021 02741 02642 04321 02621 00503 02825
                    02301 02721 00246 02341 12601 02642 02541 01284 02561 02881
                    01641...
                </u>
            </body>
        </text>
        <text id='tlsg01tagged'>
            <body>
                <u>
                  <anchor id="tlsg01tagged0000" />
                  <w type="UH" lemma="ehm">ehm</w>
                  <w type="UH" lemma="well">well</w>
                  <w type="VM" lemma="could">could</w>
                  <w type="PPY" lemma="you">you</w>
```

```
                <w type="VVI" lemma="tell">tell</w>
                <w type="PPIO2" lemma="we">us</w>
                <w type="MD" lemma="first">first</w>
                <w type="IO" lemma="of">of</w>
                <w type="DB" lemma="all">all</w>
                <w type="CS" lemma="where">where</w>
                <w type="PPY" lemma="you">you</w>
                <w type="VABDR" lemma="be">were</w>
                <w type="VVN" lemma="born">born</w>
                <w type="RR" lemma="please">please</w>
                <w type="CS" lemma="where">where</w>
                <w type="PPY" lemma="you">you</w>
                <w type="VVN" lemma="born">born</w>
                <w type="II" lemma="in">in</w>
                <w type="NP1" lemma="gateshead">gateshead</w>
            </u>
              ….Remainder of the part-of-speech tagged representation of the audio
        </body>
    </text>
  </group>
</text>
```

Figure 7: Excerpt of a NECTE interview

The *<text id="tlsg01audio">* subtree contains a reference to the audio file, which obviously cannot be shown in excerpt here; *<text id="tlsg01necteortho">* contains a short extract of plain-text audio transcription; *<text id='tlsg01phonetic'>* contains numerical codes representing phonetic segments in the transcription scheme, details of which are given at the NECTE website and, finally, *<text id='tlsg01tagged'>* contains the corresponding part-of-speech tagged representation. There are various other tags, such as *<anchor id="tlsg01audio0000"/>* which time-align the audio with the corresponding other types of representation (all of which are also available at http://www.ncl.ac.uk/necte).

The multiplicity of XML tags makes NECTE very difficult to read since a plain text option is currently not available.[5] This should, however, not be a problem in principle or

[5] Future upgrades of the resource to enable this type of download are planned.

in practice. TEI-conformant corpora are not intended for direct human inspection but rather for use with XML-aware application software like *Xaira* (http://www.oucs.ox.ac.uk/rts/xaira/), which interprets the tags and uses them in whatever analysis is specified by the user and in content presentation to the user in accessible formats.

2.4 Dissemination

An important NECTE initiative as regards long-term sustainability, tools for analysis, interoperability and dissemination is the project team's involvement in a new award funded by JISC known as ENhancing RepOsitories for Language and LitErature Researchers (ENROLLER - http://www.ahessc.ac.uk/node/383). This project is a collaboration between NECTE at Newcastle University and the National e-Science Centre (NeSC), University of Glasgow STELLA Project, Oxford University Press (OUP) and Scottish Language Dictionaries Limited (SLD). It aims to address a need for corpus researchers who currently deal with distributed, non-interoperable data repositories that are often license protected. The project will demonstrate that secure access to distributed data resources with targeted analysis and collaboration tools can be delivered in a unified framework producing a greatly enhanced research repository for phonologists as well as others more widely in the arts, humanities and social sciences.

3.0 References

Allen, W.H.A., J.C. Beal, K.P. Corrigan, W. Maguire & H.L. Moisl. 2005. 'The Newcastle Electronic Corpus of Tyneside English: Future proofing a neglected local resource',

Electronic Proceedings of *Cultural Landscapes in the 21st Century,* a Forum UNESCO - University and Heritage International Seminar: **http://www.ncl.ac.uk/unescolandscapes/english/papers.php**.

Allen, W., J.C. Beal, K.P. Corrigan, W. Maguire & H.L. Moisl. 2007. 'A linguistic time capsule: the Newcastle Electronic Corpus of Tyneside English', in: Beal, J.C., K.P. Corrigan & H.L Moisl (eds.), *Creating and Digitizing Language Corpora*, volume 2: Diachronic Databases, Basingstoke: Palgrave Macmillan. 16-48.

Anderwald, L. & B. Szmrecsanyi. 2009. 'Corpus linguistics and dialectology', in A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics. An International Handbook*, pp.1126–1139 (Series: Handbücher zur Sprache und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science). Berlin/New York: Mouton de Gruyter.

Beal, J.C. 2009. 'Creating corpora from spoken legacy materials: variation and change meet corpus linguistics', in A. Renouf & A. Kehoe (eds.) *Corpus Linguistics: Refinements and Reassessments,* Special Issue of *Language and Computers* 15: 33-47. Rodopi.

Beal, J.C. & K.P. Corrigan. 2007. ' 'Time and Tyne': A corpus-based study of variation and change in relativization stategies in Tyneside English', in S. Elspass et al. (eds.) *Language History from Below-Linguistic Variation in the Germanic Languages from 1700–2000: Proceedings,* pp.99-114. Berlin/New York: Walter de Gruyter

Beal, J.C. and K.P. Corrigan, K.P. 2009. 'The Impact of Nineteenth Century Celtic English migrations on contemporary Northern Englishes: Tyneside and Sheffield Compared', in Paulosto, H. and Penttilä, E. (eds.) *Language Contacts Meets English Dialects: Festchrift for Markku Filppula on this 60th Birthday,* pp.231-255. Newcastle: Cambridge Scholars Publishing.

Beal, J.C., K.P. Corrigan, N. Smith & P. Rayson. 2007. 'Writing the vernacular:

Transcribing and tagging the Newcastle Electronic Corpus of Tyneside English, *Studies in Variation, Contact and Change*, 1 http://www.helsinki.fi/varieng/journal/volumes/01/beal_et_al.

Docherty, G. & P. Foulkes. 1999. 'Derby and Newcastle: instrumental phonetics and variationist studies', in P. Foulkes & G. Docherty (eds.) *Urban Voices: Accent Studies in the British Isles,* pp.47-71. London: Arnold.

Heslop, R.O. (1893-1894) *Northumberland Words.* London: English Dialect Society.

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

Kerswill, P. & S. Wright. 1990. 'The validity of phonetic transcription: Limitations of a sociolinguistic research tool', *Language Variation and Change* 2:225-275.

Kretzschmar, W.A., J.C. Beal, J. Anderson, K.P. Corrigan, L. Opas-Hänninen & B. Plichta. 2006. 'Collaboration on Corpora for Regional and Social Analysis', *Journal of English Linguistics*, 34, 3: 172-205.

Local, J. K. 1982. 'How many vowels in a vowel?', *Journal of Child Language* 10:449-453.

Local, J. K., J. Kelly & W. H. G. Wells. 1986. 'Towards a phonology of conversation turn-taking in Tyneside', *Journal of Linguistics* 22:411-437.

Maguire, W. 2007. *What is a merger, and can it be reversed? The origin, status and reversal of the 'NURSE-NORTH Merger' in Tyneside English.* Unpublished PhD dissertation, Newcastle University.

Milroy, J., L. Milroy, S. Hartley, D. Walshaw. 1994. 'Glottal stops and Tyneside glottalisation: competing patterns of variation and change in British English', *Language Variation and Change* 6: 327-357.

Milroy, J., L. Milroy & 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, & 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.L., W. Maguire & W. Allen. 2006. 'Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English', in F. Hinskens (ed.) *Language Variation. European Perspectives,* pp.127-141. Amsterdam: John Benjamins.

Moisl, H. & W. Maguire. 2008. 'Identifying the main determinants of phonetic variation in the Newcastle Electronic Corpus of Tyneside English', *Journal of Quantitative Linguistics* 15: 46-69.

Pellowe, J., & V. Jones. 1978. 'On intonational variety in Tyneside speech', in P. Trudgill (ed.) *Sociolinguistic Patterns in British English*, pp.101-121. London: Arnold.

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

Preston, D.R. 1985. 'The Li'l Abner syndrome: Written representations of speech', *American Speech* 60(4): 328-336.

Preston, D.R. 2000. '*Mowr and mowr bayud spellin*: Confessions of a sociolinguist', *Journal of Sociolinguistics* 4(4): 614-621.

Strang, B. M. H. 1968. 'The Tyneside Linguistic Survey', *Zeitschrift für Mundartforschung, NF 4* (Verhandlungen des Zweiten Internationalen Dialecktologenkongresses), pp.788-794. Wiesbaden: Franz Steiner Verlag.

Tagliamonte, S. 2007. 'Representing real language: Consistency, trade-offs and thinking ahead', in Beal, J.C., K.P. Corrigan and H.L Moisl (eds.), *Creating and Digitizing*

Language Corpora, volume 1: Synchronic Databases, pp.205-240. Basingstoke: Palgrave Macmillan.

Watt, D. & L. Milroy. 1999. Patterns of variation in three Newcastle vowels: is this dialect levelling?, in Foulkes, P. & G. Docherty (eds.), pp.25-46.

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: 44-63.

Wells, J.C. 1982a. *Accents of English 1: An Introduction.* Cambridge: Cambridge University Press.

Wells, J.C. 1982b. *Accents of English 2: The British Isles.* Cambridge: Cambridge University Press.