HANDBOOK

TO THE NEWCASTLE ELECTRONIC

CORPUS OF TYNESIDE ENGISH 2

(NECTE2)

http://research.ncl.ac.uk/necte2

INTRODUCTION TO NECTE2:

NECTE2 is a corpus compiled by academics at Newcastle University comprising a wealth of recent Tyneside speech and transcript data such as sociolinguistic interviews, transcriptions, word lists and reading passages. The project was created as a follow-up to the NECTE corpus (which can be found along with full details at www.ncl.ac.uk/necte). NECTE2 adds more recent speech data and therefore another time slice to the NECTE corpus.

Currently, the corpus holds data collected between 2007 and 2009. However, a file upload page is also being developed that will allow NECTE2 to be added to in future years.

The original NECTE amalgamates two corpora, the Tyneside Linguistic Survey (TLS) corpus, with interviews conducted in the late 1960s and early 1970s, and the Phonological Variation and Change in Contemporary Spoken English (PVC) project, with interviews conducted between 1991 and 1994. Together with NECTE2 the corpora already cover an impressive time depth. As the table below illustrates, the three corpora together allow researchers to trace Tyneside speech over 50 years from speakers potentially born almost 100 years apart.

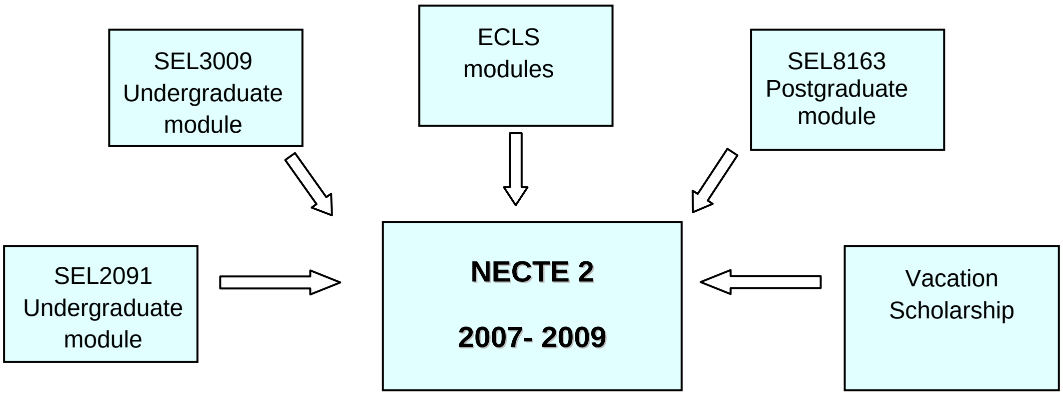
TABLE 1: EARLIEST POSSIBLE AND LATEST POSSIBLE BIRTHDATES FOR THE SPEAKERS IN EACH CORPUS, DIVIDED BY AGE GROUP

Corpus and years collected	**Younger speaker birthdates (age 17-34)**	**Older speaker birthdates (age 35+)**
TLS 1965-1970	1925- 1968	1895- 1934
PVC 1991-1994	1954- 1977	1911- 1953
NECTE2 currently 2007-2009	1967- 1990	1923- 1966

The NECTE2 corpus is intended both for researchers and anyone with an interest in the Tyneside accent and dialect. To gain access to the corpus please send an e-mail request to Dr. Isabelle Buchstaller at i.buchstaller@ncl.ac.uk .

NECTE2 SOURCES

The data in NECTE2 come from a variety of different sources:



SEL3009 Undergraduate module
ECLS modules
SEL8163 Postgraduate module
SEL2091 Undergraduate module
NECTE 2
2007- 2009
Vacation Scholarship

These sources are briefly explained below:

Newcastle University Modules SEL3009, SEL2091 and SEL8163: Most of the data in NECTE2 comes from projects undertaken by students on undergraduate and postgraduate modules in the School of English at Newcastle University. Students were asked to perform an informal sociolinguistic interview, recording roughly an hour of spontaneous speech from a dyad of closely acquainted Tyneside speakers. Use of dyads was aimed at minimising observer's paradox and lessening the need for the interviewer to interject. Dyads were also matched as closely as possible for age and social class to lessen any effects of convergence or divergence between the speakers.

Interviews from these modules also have the advantage that, because each student recorded their own interview, there was a large number of fieldworkers who were often able to select people they knew to participate. This resulted in a number of interviews where informants knew not only each other but the fieldworker as well, leading to some very naturalistic speech data.

Alongside these interviews informants were normally also asked to read a reading passage and a list of words and phrases. Copies of the reading passages (*Comma Gets a Cure* and *The Rainbow* Passage) and word list used can be found in the handbook appendix.

Vacation Scholarship: This data was collected by students on a summer scholarship project at Newcastle University. The interviews were around an hour and again focused on participant dyads matched for sex and region (Newcastle, Gateshead or Sunderland). Participants were all working class and aged 65+. Vacation scholarship interviews do not come with word lists or reading passages.

ECLS: The ECLS files were produced by students on modules SPE1024 and SPE8103 from the school of Education, Communication and Language Sciences (ECLS) at Newcastle University. Interviews come from a project where each student chose a single speaker from their hometown to interview and record. Only those speakers that came from Tyneside are included in the NECTE2 corpus.

ECLS interviews normally last around 20-30 minutes but only the first 10 minutes of each interview is transcribed. Each participant was also asked to read the *Comma Gets a Cure* and *The Rainbow Passage* reading passages.

FILE LABELLING:

All files in NECTE2 follow a labelling system for ease of identification and use. The systems for all of the different sources are detailed below.

- Newcastle University SELLL Modules (SEL2091, SEL8163, SEL3009) -

Transcripts

Each filename contains the following information:

1) 'T'- to represent a transcript

2) The module number to which the transcript was attached

3) The year in which the interview was conducted

4) A number matching the transcript to any corresponding files

Thus, 'T_3009_08_1' represents a transcript from SEL3009, transcribed in 2008 and matching any sound files from the same module with filenames ending in the number 1.

In later years, students were asked to submit some more files alongside the transcripts. Thus, some modules will include a 'supplementary materials' folder. Files within this folder begin with the same label as that shown above (to match them with the relevant transcript) but will have one of the following labels attached to the end:

- **IS** = Interview Schedule File
- **DF1** = Informant 1 demographic file
- **DF2** = Informant 2 demographic file
- **DFI** = Interviewer demographic file
- **LAF** = Lexical appendix file

Thus, for example, 'T_3009_08_1_DF1' would be the demographic file for Informant 1 matching the transcript detailed above.

Sound files

Each filename contains the following information:

1) 'S' to represent a sound file

2) The module to which the sound file was attached

3) The year in which the sound file was recorded

4) A number matching the sound file to any others from the same interview and to the corresponding transcript

5) Information as to what the file contains. This will be one of the following labels:

- **EDI** = Entire Dyadic Interview. These files include the dyadic interview plus word lists and reading passages from both informants.
- **DI** = Dyadic Interview. These files contain only the sound from the dyadic interview. Corresponding reading passages and word lists will be found in the 'supplementary materials' folder.
- **RP1/2** = Reading Passage. The number corresponds to the number of the informant detailed in the corresponding transcript. Thus, 'RP1' would be read by the speaker named 'Informant 1' in the relevant transcript.
- **WL1/2** = Word list. The number corresponds to the relevant informant as detailed above.

In some cases, word list and reading passage files are combined. Thus, a file named 'WL1+RP1' would be the word list and reading passage for Informant 1. Files named either 'WL1&2+RP1&2' or 'WL+RP' are the reading passages and word lists for both informants.

- Vacation Scholarship Files -

These files are named slightly differently to the others, and do not include word lists and reading passages. Their filenames include the following information:

1) 'T' or 'S' to represent either a transcript or a sound file

2) 'VS' to show that the files were collected during a Vacation Scholarship project

3) The year in which the files were made

4) Information about the regionality and sex of the participants. This will be one of the following:

- **NW** – a dyad of women from Newcastle
- **NM** – a dyad of men from Newcastle
- **SW** – Sunderland women
- **SM** – Sunderland men
- **GW** – Gateshead women
- **GM** – Gateshead men

These labels will be found at the end of both the transcript and corresponding sound file.

– ECLS Files –

Transcripts

Each filename contains the following information:

1) 'T'- to represent a text file

2) 'ECLS' showing that the files come from the ECLS school

3) The year in which the interview was conducted

4) A number matching the transcript to any corresponding files

Thus, a sample filename would be **T_ECLS_2009_5**.

Text files named as above are the interview transcriptions. Those named as above with the addition of '_notes' on the end of the filename are additional notes about the interview in question made by the data collectors.

Sound files

Each filename contains the following information:

1) 'S' to represent a sound file

2) 'ECLS' showing that the files come from the ECLS [module]

3) The year in which the interview was conducted

4) A number matching the sound file to any corresponding files

5) A letter indicating what the recording is:

R = *The **R**ainbow Passage* reading passage
C = ***C**omma Gets a Cure* reading passage
I = Sociolinguistic **I**nterview

Thus, a sample filename would be **S_ECLS_2009_4_C**.

NECTE2 SOCIAL MATRICES:

Below are a set of social matrices showing the social characteristics of the speakers in the NECTE2 corpus. Matrices are organised by file group, then by year and finally there is a complete matrix for the corpus as a whole.

In the matrices 'young' is defined as under 35 and 'old' is defined as 35+. Exact ages of speakers are listed with the transcripts themselves.

Please note that because social information was missing from a small minority of files, figures in the matrices are not exact as they only include speakers about whom complete social information was provided.

SOCIAL MATRICES BY FILE GROUP:

NECTE2 08:

Vacation Scholarship files [2007]:

	F		M	
	WC	MC	WC	MC
Young	0	0	0	0
Old	6	0	6	0

SEL2091 [2007]:

	F		M	
	WC	MC	WC	MC
Young	12	13	9	24
Old	5	2	1	0

SEL8163 [2007]:

	F		M	
	WC	MC	WC	MC
Young	0	5	1	3
Old	0	3	2	2

SEL3009 [2008]:

	F		M	
	WC	**MC**	**WC**	**MC**
Young	0	0	0	0
Old	3	4	1	2

NECTE2 09:

SEL2091 [2009]:

	F		M	
	WC	**MC**	**WC**	**MC**
Young	6	23	15	25
Old	7	5	3	1

SEL3009 [2009]:

	F		M	
	WC	**MC**	**WC**	**MC**
Young	0	0	0	0
Old	2	3	1	2

SEL8163 [2008]

	F		M	
	WC	MC	WC	MC
Young	0	1	0	1
Old	0	0	0	2

ECLS

(NB: Because these files came from a mixture of years between 2007-2009 they have been omitted from the matrices that are organised by year)

	F		M	
	WC	MC	WC	MC
Young	0	5	0	2
Old	0	0	0	0

SOCIAL MATRICES BY YEAR DATA COLLECTED:

2007:

	F		M	
	WC	MC	WC	MC
Young	12	18	10	27
Old	11	5	9	2

2008:

	F		M	
	WC	MC	WC	MC
Young	0	1	0	1
Old	3	4	1	4

2009:

	F		M	
	WC	MC	WC	MC
Young	6	23	15	25
Old	9	8	4	3

SOCIAL MATRIX FOR THE WHOLE OF NECTE2, 2007-2009 (including ECLS speakers):

	F		M	
	WC	MC	WC	MC
Young	18	47	25	55
Old	23	17	14	9

ETHICS:

Ethical practice was of paramount importance when compiling the NECTE2 corpus. Before the interviews began all participants were asked to read a description of the assignment in which they were participating and sign a consent form detailing where the materials were stored and how they could be used [see appendix]. All participants were fully aware that they were being recorded and consented to this.

Participants were also allowed to choose a pseudonym and were told that their real names would not be used. Accordingly, in all of the files any records of real names have been deleted and any mentions throughout the interview transcripts have been blanked and replaced with 'XXX'.

The NECTE2 researchers have done their utmost to preserve the anonymity of all participants. In some cases, however, this has not been entirely possible; for example regarding mentions of names in sound files. We would therefore ask that you exercise caution and be sure to fully anonymise participants in any publications that make use of the corpus.

TRANSCRIPTION CONVENTIONS:

Below are some of the most common transcription conventions used in the NECTE2 interviews. A more comprehensive list can be found in the Orthographic Transcription Protocol (OTP) [see appendix].

{ = Interruption and overlap

E.g. [08-09/0/JD] @ Janine it's your turn n {ow. @
[08-09/'N'(CJ)595] {Did you enjoy that?

(N) = Pause (N= length in seconds)

... = Medium pause

.. = Short pause

:::::: = Syllable lengthening, according to its duration

@ = Laughter

† = Cough

<XX> = Indecipherable passage

< > = Unclear passage. Words within brackets represent transcriber's approximation of what was said.

E.g. [07-08/T/MP/158] <That like> used to be next to The-Gate

(( )) = Transcriber's comment

NECTE2 Studies

The following list comprises the studies to date that have made use of the NECTE2 corpus. The list should help to give an idea of some of the kinds of work that NECTE2 can be suitable for.

Barnfield, K. 2008. *That's canny Geordie, man: The patterning of intensifiers in Tyneside speech.* Unpublished undergraduate paper, University of Newcastle upon Tyne.

Barnfield, K. 2009. *Intensifiers in Tyneside: Using Corpora to Observe Trends in Real-Time.* Unpublished undergraduate dissertation, University of Newcastle upon Tyne.

Barnfield, K. and I. Buchstaller. 2009*. The Newcastle Corpus of Tyneside speech 2: A novel resource for investigating longitudinal change.* Talk presented at the UK Language Variation and Change conference 7. Newcastle University, September 1-3, 2009.

Barnfield, K. and I. Buchstaller. *Intensifiers in Tyneside: Longitudinal Developments and New Trends.* Journal of Sociolinguistics, under review.

Barnfield, K. and I. Buchstaller. 2009*. Intensifiers in Tyneside: Longitudinal Developments and New Trends.* Talk to be presented at the 38th New Ways in Analyzing Variation Conference. Ottawa, October 22-25th, 2009.

Buchstaller, Isabelle and Karen Corrigan 2009. *Introducing the Diachronic Corpus of Tyneside Speech (DECTE): Methods, Data and Applications.* Paper to be presented at the 38th New Ways in Analyzing Variation Conference. Ottawa, October 22-25th, 2009.

Coverdale, D. 2008*. A layered approach to morpho-syntactic variability in Northern Irish English.* Unpublished undergraduate paper, University of Newcastle upon Tyne.