Creating and Digitizing Language Corpora for Research and Public Engagement: The Diachronic Electronic Corpus of Tyneside English (DECTE) and the Talk of the Toon

Karen Corrigan (Professor of Linguistics and English Language, Newcastle University, UK)

<k.p.corrigan@ncl.ac.uk>

The North East of England has a rich cultural heritage, not least in relation to local language. The Diachronic Electronic Corpus of Tyneside English (DECTE; Corrigan et al. 2010-2012) is a project based in the School of English Literature, Language and Linguistics (SELLL) at Newcastle University in the UK which seeks to preserve, record and capitalize on this linguistic heritage, focusing not only on research, but also research-led teaching, outreach and public engagement.

DECTE is a corpus of sociolinguistic interviews with North East residents that builds on legacy materials collected by earlier projects in the 1970s and 1990s. These were first amalgamated in the AHRC-funded Newcastle Electronic Corpus of Tyneside English (Corrigan et al. 2001-2005), and are now being augmented with further interviews collected by researchers and students at Newcastle since 2007. As of 2017, DECTE contains 718 interviews, capturing more than 1200 local speakers in over five million words of text and 450 hours of audio. In terms of birthdates, it spans over 100 years, with the oldest speaker born in 1891 and the youngest in 1995. The dataset’s coverage is therefore unrivalled by any similar UK regional dialect archive, and is matched internationally only by the Origins of New Zealand English corpus. DECTE has been used to address research questions in various subfields of English linguistics, such as phonetics/phonology (e.g. Corrigan 2012, Corrigan et al. 2013, Moisl 2015, Moisl & Jones 2005, Moisl et al. 2006, Moisl & Maguire 2008), morphosyntax (Buchstaller & Corrigan 2015, Buchstaller 2016, Childs et al. 2015, Corrigan et al. 2013, Fehringer & Corrigan 2015a/b/c), and discourse (Barnfield & Buchstaller 2010, Buchstaller 2011, 2015, 2016). Other publications have focused on the state-of-the-art architecture of the corpus, outlining tried-and-tested methods and examples of good practice for the

In parallel with linguistic research and corpus development, a focus on wider engagement and impact has been a core concern of the DECTE project (Mearns et al. 2016). The current phase of the corpus arose from an ongoing teaching and learning initiative. This aims to improve students’ understanding of sociolinguistic fieldwork methods, and to provide a context in which they can develop transferable skills related to interview techniques, transcription, data processing and analysis. We have also used DECTE in: (a) sessions for primary, secondary and A-Level students, covering diverse language-related curriculum topics; (b) CPD events for A-Level English teachers; (c) public lectures; (d) booklets for sale at local museums; and (e) an interactive, public-facing website, The Talk of the Toon (http://research.ncl.ac.uk/decte/toon).

This workshop will draw on the experiences of the DECTE team, our collaborators and students to discuss some of the challenges we have faced, in relation to four broad themes:

(1) **Corpus Construction**: What are the gold standards for conducting an ethically sound sociolinguistic interview and what best practices are there for transcribing and processing this kind of linguistic data?

(2) **Sustainability**: How can we ensure that corpus resources developed in academic environments, supported by public funding bodies, are ‘future-proofed’ so that the investment in them is not wasted?

(3) **Research Value**: How can we ensure that the research effort to create DECTE is capitalized on by us and other researchers in answering important research questions?

(4) **Relevance**: How can we take the results of our academic work beyond Higher Education to schools, museums and the public, in order to achieve the widest possible impact?

**References:**


