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 development of other similar projects (Allen et al. 2007, Beal & Corrigan 2013, Beal et al. 2014, Corrigan 2017, Kretzschmar et al. 2006, Mearns et al. 2016).

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) R*esearch 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:

Allen, W., Beal, J.C., Corrigan, K.P., Moisl, H. and Maguire, W. 2007. 'The *Newcastle Electronic Corpus of Tyneside English'*, in Beal, J.C., Corrigan, K.P. and Moisl, H. (eds.) *Creating and Digitizing Language Corpora: Vol. 2, Diachronic Databases,* pp.16-48*.* Houndsmills: Palgrave Macmillan.

Barnfield, K. and Buchstaller, I. 2010. 'Intensifiers on Tyneside: Longitudinal developments and new trends', *English World-Wide* 31: 252-287.

Beal, J.C. and Corrigan, K.P. 2013. 'Working with unconventional existing data resources', in Childs, B., Mallinso, C., van Herk, G. (eds.) *Data Collection in Sociolinguistics: Methods and Applications,* pp.213-216. London: Taylor & Francis.

Beal, J.C., Corrigan, K.P., Mearns, A.J. and Moisl, H.L. 2014. 'The Diachronic Electronic Corpus of Tyneside English: Annotation and dissemination practices', in Durand, J., Gut, U. and Kristoffersen, G. (eds.) *The Oxford Handbook of Corpus Phonology,* pp.517-533*.* Oxford: Oxford University Press.

Buchstaller, I. 2011. 'Quotations across the generations: A multivariate analysis of speech and thought introducers across 5 generations of Tyneside speakers', *Corpus Linguistics and Linguistic Theory* (Special issue on Corpus Linguistics and Sociolinguistics, Tylor Kendall and Gerard van Herck eds.), pp.59-92.

Buchstaller, I. 2015. 'Exploring linguistic malleability across the lifespan: Age-specific patterns in quotative use', *Language in Society* 44(4): 457-496.

Buchstaller, I. 2016. 'Investigating the effect of socio-cognitive salience and speaker-based factors in morpho-syntactic life-span change', *Journal of English Linguistics* 44(3): 199-229.

Buchstaller, I. and Corrigan, K.P. 2015. '"That's bad grammar again isn't it?": (Morpho)-syntactic features of Northern English', in Hickey, R. (ed.) *Researching Northern Englishes,* pp.71-98*.* Amsterdam: John Benjamins.

Childs, C., Harvey, C., Corrigan. K.P. and Tagliamonte, S.A. 2015. 'Comparative sociolinguistic insights in the evolution of negation: Selected papers from *NWAV 43. Penn Working Papers in Linguistics* 21: 2. <http://repository.upenn.edu/pwpl/vol21/iss2/4/>

Corrigan, K.P. 2012. 'GOAT vowel variants in the Diachronic Electronic Corpus of Tyneside English', in Nevalainen, T. and Traugott, E. (eds.) *Rethinking Approaches to the History of English,* pp.90-93. Oxford: Oxford University Press.

Corrigan, K.P. 2017. 'Corpora for regional and social analysis', in Montgomery, C. and Moore, E. (eds.) *Language and a Sense of Place: Studies in Language and Region,* pp.107-127. Cambridge: Cambridge University Press.

Corrigan, K.P., Mearns, A.J. and Moisl, H. 2013. 'Feature-based versus aggregate analyses of the DECTE corpus: Phonological and morphological variability in Tyneside English', in Szmrecsanyi, B. and Wälchi, B. (eds.) *Aggregating Dialectology, Typology and Register Analysis,* pp.113-149. Berlin/Boston: Walter de Gruyter Gmbh.

Corrigan, K.P., Moisl, H.L. and Beal, J.C. 2000-2005. *A Linguistic 'Time-Capsule': The Newcastle Electronic Corpus of Tyneside English.* <http://research.ncl.ac.uk/necte/>.

Corrigan, K.P., Buchstaller, I., Mearns, A.J. and Moisl, H. 2010-2012. *A Linguistic 'Time- Capsule' for the Google Generation: The Diachronic Electronic Corpus of Tyneside English.* <http://research.ncl.ac.uk/decte/>.

Fehringer, C. and Corrigan, K.P. 2015a. 'The rise of the *going to* future in Tyneside English: evidence for further grammaticalisation', *English World Wide,* 36(2): 198-227.

Fehringer, C. and Corrigan, K.P. 2015b. '"The Geordie accent has a bit of a bad reputation": Internal and External Constraints on Stative Possession in the Tyneside English of the 21st Century', *English Today,* 31 (2): 38-50*.*

Fehringer, C. and Corrigan, K.P. 2015c. '"You've got to sort of eh hoy the Geordie out": Modals of obligation and necessity in 50 years of Tyneside English', *English Language and Linguistics* 19 (2): 355-381.

Kretzschmar, W., Anderson, J., Beal, J.C., Corrigan, K.P., Opas-Hänninen, L. and Plichta, B. 2006. 'Collaboration on corpora for regional and social analysis' *Journal of English Linguistics,* 34: 172-205.

Mearns, A.J., Corrigan, K.P. and Buchstaller, I. 2016. 'The Diachronic Electronic Corpus of Tyneside English and The Talk of the Toon: Issues in preservation and public engagement', in Corrigan, K.P. and Mearns, A.J. (eds.) *Creating and Digitizing Language Corpora, Volume 3: Corpora for Public Engagement.* Houndmills, Basingstoke: Palgrave-Macmillan, pp.177-210.

Moisl, H. 2015. *Cluster Analysis for Corpus Linguistics.* Berlin: De Gruyter.

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

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

Moisl, H., Maguire W. and Allen W. 2006. 'Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English', in Hinskens, F. (ed.) *Language Variation. European Perspectives,* 127-142. Amsterdam: John Benjamins.

Moisl, H. 2014. 'Statistical corpus exploitation', in Durand, J., Gut, U., Kristofferson, G. (eds.) *The Oxford Handbook of Corpus Phonology,* pp.110-132. Oxford: Oxford University Press.