Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English

Will Allen (w.h.a.allen@ncl.ac.uk)
Warren Maguire (w.n.maguire@ncl.ac.uk)
Hermann Moisl (hermann.moisl@ncl.ac.uk)

School of English Literature, Language, and Linguistics
Percy Building
University of Newcastle
Newcastle upon Tyne NE1 7RU
United Kingdom

The Newcastle Electronic Corpus of Tyneside English is a corpus of dialect speech from Tyneside in North-East England. It includes narrow phonetic transcriptions of 64 speaker interviews, which we have begun to analyze using exploratory multivariate techniques. The present paper describes the methodology and results to date.

The data set is very high-dimensional (642 phonetic segment variables) relative to the number of data items (64), and is thus very sparse. To understand the shape of the data manifold, we have used principal component analysis to reduce the dimensionality of the data space by eliminating correlations among the variables, and the result has been that up to 95% of the covariance in the original data can be restated using no more than 25 variables. Current work includes interpretation of the derived variables, application of other dimensionality reduction techniques whose output takes account of any nonlinearities in the data, and cluster analysis of the reduced-dimensionality data to classify speakers according to their phonetic profiles.