William A. Kretzschmar, Jr., University of Georgia, USA, kretzsch@uga.edu
Jean Anderson, University of Glasgow, Scotland, J.Anderson@arts.gla.ac.uk
Joan Beal, University of Sheffield, England, j.c.beal@shef.ac.uk
Karen Corrigan, University of Newcastle , England,
karen.corrigan@btinternet.com
Lisa-Lena Opas-Hänninen, University of Oulu, Finland, lisa.lena.opas-hanninen@oulu.fi
Bartek Plichta, Michigan State University, USA, plichtab@msu.edu

Collaboration on Corpora for Regional and Social Analysis

Compilers of corpora that document regional and social languages and varieties of languages must necessarily have different needs and goals, and yet we also face common problems. Thus we have an interest in collaboration, for use of computer tools ourselves and for greater ease of use by our audiences. In this paper, we set forth our intention to begin such a collaboration. We begin by exploring the parameters of our various corpora, whether historical or newly created, whether of written texts and/or of speech, whether recorded in writing or audio or video:

The American Linguistic Atlas Project, written and spoken data collected in surveys from the 1930s through the present day.
The SCOTS Project, both written and spoken texts for the languages of Scotland.
The Newcastle Electronic Corpus of Tyneside English (NECTE), data from two surveys conducted in 1969 and 1994.
The Corpus of Sheffield Usage, data from the Survey of Sheffield Usage collected in 1981.
The Finnish Arctic Language Project, video and other records from marginalized language populations in the Artic to be collected soon.

We then explore the parameters of access and analysis, whether public or private, whether for general audiences or for specialists. Finally, we assert that, despite the evident difficulties of collaboration for such disparate ends, it is indeed possible, practical, and desirable for us to apply common methods to our common problems. Further, we propose specific recommendations for what methods we should begin to apply in our work, beginning with emerging international standards for metadata for language archives (Dublin Core, OLAC), and continuing with best practices for the collection, preservation, and presentation of corpus data in our work.