

Panel on Demographic Coding for Sociolinguistic Corpus Archive Preparation

Organizers: Christopher Cieri of LDC (ccieri@ldc.upenn.edu), and Malcah Yaeger-Dror of University of Arizona and LDC (malcah@email.arizona.edu).

Abstract

This two-hour panel addresses one aspect of **Sociolinguistic Archive Preparation**. A unified archive for our field would benefit from a common and coherent set of coding conventions. Comparable demographic protocols would facilitate not only comparison of individual speech community studies but also collaboration across research groups and the accumulation of findings from individual project to reveal broader regional and national patterns and trends. This panel will contribute to that goal by discussing the importance of shared conventions for coding demographics. Each panelist will present 10 minutes of discussion on either the concept of ‘sharing metadata’ [Panel I] or a specific demographic variable which s/he feels should be more carefully determined [Panels II-III]

Panel goals are:

- 1) to document the need for more extensive and detailed demographic categories based on field experience
- 2) to sketch, both a superset of demographic categories from which individual researchers may choose to select as appropriate to their community
- 3) and a core set of demographic categories and values that should be present in all studies to permit comparability
- 4) to discuss options for publicly sharing the definition of these demographic categories and to select at least one approach for doing so in the future to promote the use of a core set of demographic categories

TIME	Moderator	Presenter	Featured Topic
10	Yaeger		I Sharing
10		Eskenazi	Sharing Metadata
10		Terry Langendoen	Sharing Metadata : GOLD
10		MacWhinney	Sharing DATA
10		...+Cieri	Discussion
	Cieri		II- III Broad Demographic Categories
10		Blake	Demographics vs. Race: “African American”
10		Otheguy	Demographics vs. Race: “Hispanic”
10		Wong/Hall-Lew	Demographics vs. Race “Asian”
10		all	Discussion
10	Yaeger	Yaeger	Demographics: “Religion”
10		Bowie	Demographics: “Age”
10		all	Discussion
			IV. Resolution.
10	Cieri		Discussant: Shen
20			All presenters

The first panel will discuss the importance of *metadata for sharable archives*. The presenters in the first half hour are all in charge of large sharable archival corpora – Christopher Cieri at Linguistics Data Consortium [<http://www.upenn ldc.edu/>; <http://projects ldc.upenn.edu/>] will act as moderator for Terry Langendoen, who has helped to develop the GOLD standard, Brian MacWhinney who developed the CHILDES & TALKBANK archives [<http://www.talkbank.org/>], and Maxine Eskenazi who has built large corpora like ‘Let’s Go!’

[<http://www.speech.cs.cmu.edu/letsgo/>]. They also communicate with other possible user groups for sociolinguistically coded data.

The presenters in the second segment are each carrying out research in communities for which the ‘ethnic’ designations common to our field are obviously inappropriate, or at best under-differentiated [cf., pewresearch.org]. Renée Blake compares ‘African American’ speakers from different backgrounds; Fought will compare Latinos from different communities; Hall-Lew and Wong compare speakers from within the broadly defined ‘Asian’ community.

Presenters in the third segment are interested in variables which are not always coded for; Bowie is carrying out ground-breaking work on variation within one religious community, and both Yaeger and Bowie are very interested in age and the importance of change in real time.

The last session will open with a discussion by Wade Shen, a project director from Lincoln Laboratories [<http://www.ll.mit.edu/mission/communications/HLT/HLT.html>], which often suggests corpora to be developed by LDC, and which makes use of large sociolinguistic corpora in their own sociophonetic research. Shen and the presenters from the earlier sessions will raise questions which will be discussed by presenters and other participants.

The recent discussion on Var-L (the sociolinguists’ online discussion forum -- VAR-L@JISMAIL.AC.UK¹) has highlighted the degree to which our field is primed for the development of larger archival corpora. We anticipate that this workshop will bring together those interested in pursuing large-scale collaborative sociolinguistic research, coordinating cross-regional and cross-group coding. The groundwork laid by this workshop is expected to provide a platform for continued discussion at a Satellite Workshop preceding the Winter LSA.

¹ Cf. particularly http://bitly.com/VAR-L_Jun11, particularly those posts between #2011 and 2048.