Introduction to the workshop

Evolving question formats for data coding conventions.¹

For the past several years the Linguistics Data Consortium at the University of Pennsylvania [LDC url: www.upenn.ldc.edu] has been collecting and archiving linguistic data for future use. At different points over these years I have made use of the LDC corpora, as well as the Montreal French Corpus, and various Talkbank corpora. I have also had the fortune [or misfortune] to be a frequent referee. In the process it has become quite clear that sociolinguists need help from those who have carried out previous archives, because we often assume that our corpora have transparently followed a shared protocol for research, and which permits them to be compared, when in actuality, there are data gathering conventions which are not shared, and are not even expressly stated so that others can determine the degree to which corpora are sharable even within the same community.

There are exceptions, primarily in Canada [Toronto-Ottawa-Montreal], where corpora follow the conventions first discussed in the work of Sankoff and Cedergren (Sankoff & Sankoff 1973), and further elaborated when the 1984 Montreal corpus required a set of rules to permit comparability (Thibault/Vincent 1989/1990). There are also comparable corpora in the UK, where even in the ‘infancy’ of the BNC, the research community’s focus was on corpus gathering that would permit archival comparisons.

In addition, PADS#94 was made up of corpora that were built as parallel corpora of data from different areas of the country so they could be shared, and there has been a noteworthy effort on the part of NYU researchers, and those who are carrying out parallel studies in other cities, to formulate protocols which will permit comparability.

HOWEVER, the bulk of recent corpora are not necessarily comparable, and there is not always a way to determine the degree to which a corpus conforms to the general rules for sociolinguistic corpora or not.

At the same time, the NSF has enacted new rules which require that we all document in a proposal how the data will be made sharable by researchers from other communities. The members of the Variationist List have also been discussing how best to gather data, so that it is later sharable. With all that in mind, Chris Cieri from LDC and organized this workshop to help address some of the difficulties which we know will crop up.

1. **DATA**: The first section will consist of speakers who already have been working with very large corpora, making them available in open archives, and even making their transcription and coding available….Brian MacWhinney ([www.talkbank.org](http://www.talkbank.org)), Gary Simons ([www.sil.org](http://www.sil.org); [www.gold….](http://www.gold….)), Maxine Eskenazi (), and Wade Shen ([www.ll.mit.edu](http://www.ll.mit.edu)) will all provide some perspective.

2. **Ethics Issues**: How do we maximize our ability to be sensitive to the ethical issues involved in setting up a project, while minimizing the amount of time on IRB genuflection/expediting the IRB process, so (student) researchers can carry out their studies? Natasha Warner from the LSA’s own ethics committee and Denise DiPersio from LDC will both make suggestions and answer questions.

3. **Metadata**: --Labov Rule #1 is:Code for any variable that might be relevant, because one cannot go back and add more distinctions later without totally disrupting the study. But what information is the minimal amount of information that we should store? How do we ask for it [given that different types of questioning will get different [and probably noncomparable] answers? How can we insure that our question formats as well as the number of choices available are the same? There seem to be several foci for coding that are often not addressed at all, much less addressed using identical question formats. Metadata which can be shared are also in their infancy. NSF has helped us to invite several speakers

¹ Thanks to both the NSF, and the LDC for the funding which has made this study possible. Thanks also to the many sociolinguists who have agreed to take part in the NSF sponsored workshop.
Introduction to the workshop

who have already coded for specific variables, and can provide perspective on how to ask appropriate questions, so that the answers can be used for coding:

a. **Demographic coding** is constantly upgraded to reflect the information on significant community subgroups, but there are still recurring problems where our coding has turned out to be too ‘coarse grained’, but most studies include
   i. birthdate,
   ii. speaker sex,
   iii. speaker age at the time of recording,
   iv. racial heritage,
   v. and some form of discussion of other sociodemographic factors.

b. **Demographics** we are less ‘on top of’ include such issues as:
   i. the regional heritage of the speaker; often studies retain an outmoded under-differentiated regional coding scheme. Since this type of coding is more dependent on the specific region one is working in, it will not be discussed here today.
   ii. While speaker sex has been consistently coded for over the last several years, evidence has been piling up that sexuality can also be a critical determinant of speaker variation, and Penny Eckert will present a background paper on the importance of sexuality as a factor in language variation…. iii. Ethnicity—specifically ethnic designations that we all tend to overgeneralize in our work—will form one section today, with Renée Blake discussing various social groups which do not consider themselves ‘African American’, or which have multiple self-identities some of which may outweigh that ethnic identity, at least some of the time; Amy Wong & Lauren Hall-Lew discuss various ‘Asian’ subgroups for which ‘Asian’ might be an inappropriate identifier, and Carmen Fought discussing the fact that ‘Latino’ is not one uniform ethnicity, but that different national, regional, racial or ethnic sub-identities are more salient in many—or even most—situations.

iv. The speakers’ educational achievements are generally found somewhere in the transcript even if they haven’t been coded for, but while earlier studies were prone to code for a rather rigid ‘socioeconomic scale’, which was explicated in the text, recent studies may not even code for how a speaker makes a living, much less the SES or ML.

v. Despite the early work of the Milroys (e.g., 1987), the work of descriptive linguists, and of social psychologists like Giles and Bourhis (1973), which has shown that the religious and political background of the speakers strongly influences social attitudes and opinions (which in turn influence speech), both political and religious persuasion are generally ignored. Even when they are discussed, information on individual religious persuasion (much less commitment to that religion) is very rarely considered in a sociolinguistic study. [When religion is coded, it is relegated to a category within ‘ethnicity’ for lack of a generally accepted ‘religion’ category. It is embarrassing to read the work of descriptive linguists, like Catherine Miller (2005) or Clive Holes (e.g., 1986), who have been describing for years the degree to which dialect features which initially appear to be regional are actually traceable to speakers’ degree of religious commitment and political attitudes, and those, in turn, influence the dialect variation, so that a more conservative part of a country will differ considerably from the more modern areas. Sociolinguists have not been coding for either of these. David Bowie will discuss questions which have been found to successfully provide this
4. **Social Attitudes:** Although Howard Giles and many other social psychologists publish work on how social attitudes influence language choice, it is only recently that sociolinguists have begun to ask questions that will permit us to code for relevant attitudes in our analyses of dialect variation. Carmen Llamas (Llamas, Watt & Johnson 2009; Watt, Llamas & Johnson 2010), Naomi Nagy (2011, 2012) and their colleagues, and Shana Poplack (e.g., 2007) have all spent years comparing ways of eliciting information about social attitudes without contaminating the data with the ‘interviewer effect’, and Kim Noels, from the social psychology community which studies language variation, will all share some of their insights on how to ask such questions, how to preserve the coded responses to permit later comparison with the linguistic results, and what questions may be most useful.

5. **Situation:** While we generally assume that we have been coding for the social situation, and there have been many studies of ‘style’ or ‘register’, quite often individual corpora totally ignore situation, do not code for it, even superficially, and it is not always automatically determinable from a transcript, even when there is one. This is all the more embarrassing given not only Labov’s work, but that of Hymes (1964), and Giles (as early as Giles & Powesland 1975), as well as the more recent work which is often cited (Eckert & Rickford 2001; Bell 1984, 2001; Coupland 2007….), which all point out that [among other things] one aspect of the social situation is actually dependent on the speakers’ attitudes toward their interlocutors’ social group memberships. Sali Tagliamonte will be presenting some coding conventions which she feels have served her well for the analysis of interview data, and which we can follow, to carry out more appropriate coding of situation, while John Rickford will also consider situational features which are needed for data from other social settings….

6. **Discussion & Conclusions:** The last section will be devoted to a generalized discussion, so those about to go carry out fieldwork can ask questions, and those who have been carrying out fieldwork, and who have perhaps discovered more efficient ways to code and archive can do so. All the speakers will be available, as will others, like Tyler Kendall, who have also been involved in sociolinguistic archiving of data both here and in Europe.

**References**


Introduction to the workshop


___&___ 2010b. Second generation West Indian Americans and English in NYC. *English Today* 103(26#3) 35–43.


___ & A. Black. 2010. SDC: The Spoken Dialog Challenge, invited presentation at SIGDIAL.


Introduction to the workshop


__ & A. Wong 2011. Chinese-Americans in NYC and SF. For upcoming LSA.


Holes, Clive. 1986 The social motivation for convergence in three Arabic dialects. IJSL 61: 33-51


Koops, Christian 2010. /u/-Fronting is not Monolithic: Two Types of Fronted /u/ in Houston Anglos. PWPL 16. 2: Article 14


Rickford, John in press….

Introduction to the workshop


Tagliamonte, S. 2011a Peaks and trends in linguistic change. Penn colloquium.


Wong, Amy 2007. Qualifying paper on field work for study of Chinese Americans in New York City. NYU.


