Linguistic Society of America Annual Meeting
Satellite Workshop for Sociolinguistic Archival Preparation
January 4-5, 2012, Portland, Oregon

Christopher Cieri, Position Paper

Motivations

All quantitative sociolinguistics is inherently corpus based. Whether or not the practitioners explicitly set out to create a corpus as a by-product of their analysis, they generally sample and select subjects and collected and annotate (code) data for a specific purpose or else analyze data that was so collected. Any research that creates or exploits corpora faces a number of issues from design, planning, data acquisition, processing and formatting, metadata, annotation and intellectual property and human subject protection. From among these many issues, the specific subset of metadata and the human subjects protection needed in order to acquire that metadata are timely and worthy of attention.

Herein, data refers to records of observations of linguistic practice then metadata refers to information about how, from whom and under what circumstances those records were collected, in short the subset of independent variable that are related to subjects and interview sessions. Notwithstanding our focus on metadata and human subjects protections, many of the issues discussed here will prove relevant to the annotation of linguistic data as well. Although we focus on metadata issues for sociolinguistic archive preparation, it should be noted that workshop topics affect many other activities. In addition to the archive builders, those who would collect data to share or just use for their own analyses, those who would use data collected by others or compare their results to those of other studies and those who would compare the results of multiple prior studies to find broader trends must also face issues such as consistency in practice.

Ideally coding practice evolves to become, in order:

- Understood: Multiple test runs are sometimes necessary to identify all relevant variable, all their values, and the special cases. As coding practice changes to accommodate newly found information, it is necessary to re-code (or else accept inconsistency between data coded before and after the change possibly reducing the volume of the reliable data or the variables that can be used to condition analysis).
- Documented: The documentation of coding practice permits a single researcher the ability to remain internally consistent over time, especially after a break in activity. If that documentation is shared then it allows other scholars to evaluate the work along the methodological dimension, discuss differences in practice, determine which are motivated by differences in speech communities and which are the result of historical accident.
- Consistent: Consistency is practice permits studies to be compared pair-wise within and across communities. Later studies can build upon prior and panel

studies can begin to develop generalizations not possible with only single community studies. In addition teams of annotators can work together to produce larger volumes of data.

- Standardized: with the adoption of a public standard comes broader consistency in community practice which relieves the individual of some decision making and simplifies the training of future generations.

It is worth noting though that the adoption of standards can tend to reduce discussion or else focus it on a single aspect of practice. The premature adoption of standards can lead to practice that limits progress. The ISO 639 family of language codes seems to have had the effect of focusing discussion on how to efficiently codify the names of a subset of the world languages. ISO 639-2 used 2 letter codes that can index no more than 26*26=676 languages. ISO 639-3 uses three letter codes to increase the scope to encode all the ~7000 languages recognized by Ethnologue though some of the codes are less mnemonic than others. In the meantime, while organizations worked to become ISO 639-3 compliant, (proposed) standards to encode not only the language but also the location where it is spoken have received less attention. Of course any sociolinguist immediately recognizes that these early language codes are of limited use to their work which needs to maintains distinction among linguistic varieties that are typically given the same language code or language:location code pair. ISO 639 code standard were published at least as early as 2001 but the versions that have begun to take linguistic varieties seriously did not appear until around 2009.

Metadata

We propose to proceed methodically in our discussion of sociolinguistic metadata. We begin in this workshop by trying to understand the variation in practice among different leading scholars in the encoding of demographic and attitudinal metadata frequently as self-reported by the subject during enrollment or interview and situational metadata as recorded by the field worker during interview or after analysis. Practice varies along multiple dimensions including:

- metadata categories (variables coded),
- the possible values of those categories and
- the forms used to elicit the metadata from subjects (questions asked).

Each of these dimensions interacts with the others. For example the form of the questions affects the values elicited.

Our current goal is to discuss the many ways in which we elicit and encode metadata categories and values in order to determine which subset can be standardized and which must remain community specific. Even when we expect variation by community we seek to build a set of suggested practices that serve as a point of departure for new efforts so that new generations of field workers can build upon the experiences of their predecessors. A desired outcome of this discussion is a reference list of metadata categories, potential values and questions used to elicit such metadata from human subjects. The list would be somehow ordered or marked to distinguish those likely to be of general use from those likely to vary by community. Ideally, this material would be maintained and edited by the community and served as training material for new scholar and a model for new studies. In that way, any departure from the suggested practice would be the result of an affirmative decision rather than the result of an accident.

Human Subject Protection

The great majority of sociolinguistic research involving human subjects shares a very similar risk benefit ratio. The risks to subjects are generally no greater than those encountered in daily life except that the studies can pose a small and easily manageable risk to anonymity. Although subjects frequently enjoy the sociolinguistics interview and the opportunity to talk about themselves, the benefits are principally social. Given this low risk and potentially significant social benefit one would expect most Institutional Review Boards to readily approve projects including sociolinguistic interviews with appropriate controls are in place. In addition, one would expect anonymized versions of these data sets to be shareable. Instead what we observe anecdotally is wide variation in the reactions of local IRBs to interview data. A second goal of this workshop is to address human subject protocols within the context of the metadata we plan to discuss above and elicit subsequently. If the community is able to agree upon some shared metadata and a common approach to human subject protocols, the broad acceptability of the practice should reduce concern among IRBs as it regularizes protection of human subjects.

Technical Support for Common Practice

There are a number of technical resources that may serve as a model or as direct support in our effort to harmonize practice. The Open Language Archives Community (OLAC) has developed a Dublin Core compliant metadata standard for describing entire corpora. The ISLE Metadata Initiative (IMDI) has developed recommendations for corpus and session level metadata. Data Category Repositories such as ISOCat provide the means to store definitions of metadata categories and values in a persistent location suitable for reference. If ISOCat's principal contribution is as a container for metdata, the GOLD project provides an example of the stuff that goes in the container. Although GOLD does not define demographic, situational or attitudinal data categories it provides a model of the kind of resource that would benefit sociolinguists.