LSA Annual Meeting: Satellite Workshop for Sociolinguistic Archive Preparation
January 4-5, 2012, Portland, Oregon

Organizers
Malcah Yaeger
Christopher Cieri
Laurel Mackenzie
Brittany McLaughlin
Definitions

- **data** = recorded observation of linguistic event
  - speech, also written text, video of gesture, signing
- **annotation** = any application of human judgment adding value to data
  - transcription, coding of speech, text transcript
- **metadata** = information on from whom, under what circumstances data collected
  - speaker demographics & attitudes, situation
  - corpus level versus session level
- **relation to terms coding and variables**
Motivation: LDC Corpora for Sociolinguistics

- Malcah’s use of CallFriend queries about metadata
- The “e question” in Mixer
  - How to formulate it for a series of national studies?
  - Sociolinguistic Interviews in Mixer
    - 450 English speakers, 150 Spanish speakers * 3-4 sessions each
    - contrasted with conversational telephone speech, transcript reading
- Maxine’s request for more detail metadata in LDC corpora
- Brian’s inclusion of LDC corpora in Talkbank and efforts to include sociolinguistic data beyond SLx
Motivation: Sociolinguistic Corpora for Collaboration in HLT

- Data and Annotation for Sociolinguistics:
  - study of -t/d deletion across many prior studies, misalignment, underspecification
  - -t/d deletion study in TIMIT and Switchboard Corpora

- SLx Corpus of Classic Sociolinguistic Interviews
  - segmented, transcribed, sample annotation for >100 sociolinguistic variables, specification

- Wade’s attempt to use sociolinguistic data for language, dialect and speaker ID
Malcah originally proposed LDC lead workshop on robust metadata for sociolinguistic archives

But then we realized that the most interesting issues are very fundamental

Several kinds of issues

- perspective from those already working on shared data
- variables that are often neglected or badly formed
- (concern over) human subject protection
- infrastructure for harmonizing where possible
Unified archive would benefit from common coding

- comparable demographics facilitate
  - comparison of individual speech community studies
  - collaboration across research groups
  - accumulation of findings to reveal broader patterns and trends
Goals

- document need for more extensive/detailed categories based on field experience
- define superset of categories from which individual researchers
- define core set of categories and values that should be present in all studies to permit comparability
- discuss options for publicly sharing the definition of these 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
Evolution of Coding Practice

- Understood
- Documented
- Consistent
- Standard
Benefits

- economy
- ubiquity
- clarity
- uniqueness
- Stability

Compare to “speech community”

Why important to sociolinguistics

- fieldwork typically collected in speech communities
- goals: description of grammar cognizant of variation & change
- thus collaboration, comparison are critical
Infrastructure for Harmonizing Metadata

◆ Malcah’s Questionnaires
◆ OLAC
◆ GOLD
◆ ISOCAT
◆ Economy
### 3.4.2 Actors, Actor

- **Group:** Actor
- **Identifier:** Actor
- **Definition:** Groups information about one specific person in the session.
- **Encoding:**
  - Actor . Resource Ref *
  - Actor . Role
  - Actor . Family Social Role
  - Actor . Name +
  - Actor . Full name
  - Actor . Code
  - Actor . Language +
  - Actor . Ethnic group
  - Actor . Age
  - Actor . Sex
  - Actor . Education
  - Actor . Anonymous
  - Actor . [Contact]
  - Actor . Description *
  - Actor . Keys

**Comments:**

**Actor . Resource Ref**
- **Element:** Actor . Resource Ref
- **Identifier:** Actor . ResourceRef
- **Definition:** Reference to the resource in the session this specific actor is connected with in the specified role (Actor . Role).
- **Encoding:** string (XML IDREFS attribute).
  
  **Comments:** This attribute is only used if there can be confusion about which actor is connected to a specific resource. If "Actor . Resource" is not specified it can be assumed the actor is connected with all resources in the session.

**Actor . Role**
- **Element:** Actor . Role
- **Identifier:** Actor . Role
- **Definition:** The functional role of the person participating in the session.
- **Encoding:** Open vocabulary list 'Actor . Role' (4.6).
- **Comments:** The role is meant as a rough categorization of Actors such as: interviewer, consultant, contributor, computer etc. Also people responsible for the creation of the resources are included such as author, publisher, and sponsor. This is in contrast to the "Family Social Role" of an Actor that is used for example to describe relations amongst the contributors.
Feminine Gender (Concept)

http://purl.org/linguistics/gold/FeminineGender

Definition:
A gender property established on the basis of agreement, to which nouns may be assigned if 1) they inherently denote females. Additionally, but not necessarily, nouns may be assigned this value if 2) their formal properties (morphological or phonological) lead them to be assigned to the same agreement pattern as other nouns within the language that have female denotation. 3) they are arbitrarily assigned to the same agreement pattern as other nouns in the language that have female denotation. [Corbett 1991]

Usage Notes

2009-06-04 13:28:07

Languages may differ with regard to the point in the animacy hierarchy where sex-based gender assignment occurs (e.g., whether masculine or feminine is restricted to humans only, or includes domestic animals etc.). Typically, with lower-order animates, sex-based assignment will not play a role. As well as the label feminine, it is also possible to use an arbitrary label, such as an ArabicNumericalGender or RomanNumericalGender for this value. This may be the case in certain traditions (e.g., descriptions of Bantu and Daghestanian languages) where there is a larger number of gender values, or where the sex-based assignment system is limited to a particular portion of the animates.

Examples

In Chinook, some stems can take either masculine, feminine, or neuter prefixes, with concordant semantic impact on the word. Other stems take only one gender affix. The semantic basis for this classification is not completely transparent, but some tendencies do occur, e.g., large animals tend to be masculine while small ones tend to be feminine: qualities