



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

Organizers

Malcah Yaeger

Christopher Cieri

Laurel Mackenzie

Brittany McLaughlin



- ◆ 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

The screenshot shows the OLAC Language Resource Catalog website. The browser window title is "OLAC Language Resource Catalog" and the address bar shows "http://search.language-archives.org/search.html?". The page header includes "Participating Archives • OLAC • Delivered by the Penn Libraries" and a "Printer-Friendly Page" link. The main content area is titled "OLAC Language Resource Catalog" and features a search bar with the text "Search for language resources". Below the search bar, the results section shows "Results: Showing hits 1 - 50 out of 103961". A list of search results is displayed, including "Cree, Northern East: a language of Canada", "EV05Tsinon -- Tsinon dainta -- Organization of a Text Collection in Trumai, Aiming at its Scientific Documentation", "aer2infiel -- Personal narrative - Narración personal -- Lowland Chontal Documentation", "Muyuw Orthography", "028 Stok -- JJ-Stok-Yurakare-IMG_7728 -- The Documentation of Yurakaré", "How the birds were given their responsibilities", "Handbook of the Linguistic geography of New England", "Sierra Popoluca syllable structure", and "Konabéré: a language of Burkina Faso". On the right side, there are filters for "Currently Used Filters" (None), "Sort Results By" (Possible Sorts: Title, Id, Date), and "Narrow Results By" (Archive, Online, Subject language, Language family). The footer of the browser window shows search controls: "Find: Mixer", "Next", "Previous", "Highlight all", and "Match case".

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

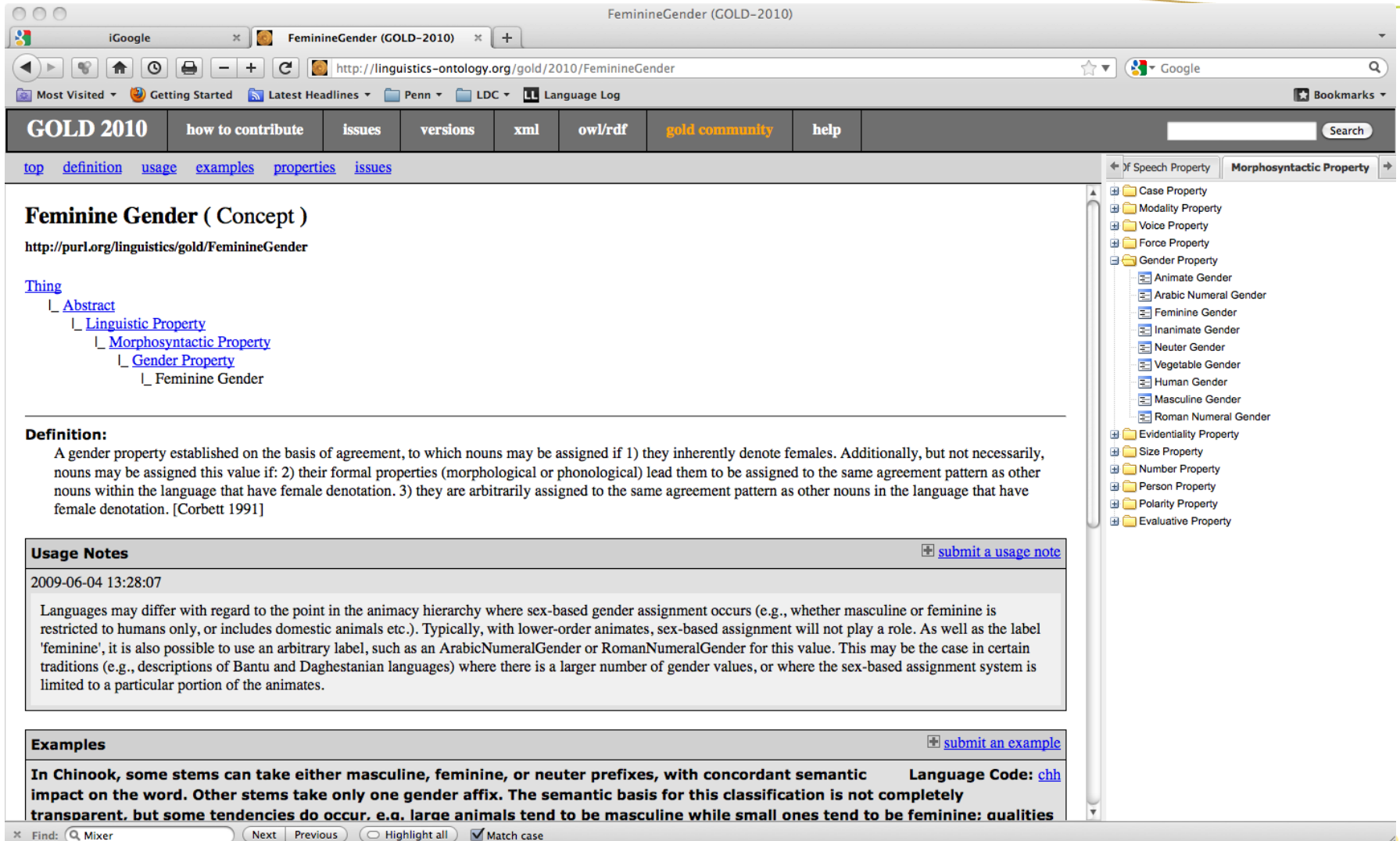
Comments:

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.



The screenshot shows a web browser window with the URL <http://linguistics-ontology.org/gold/2010/FeminineGender>. The page title is "Feminine Gender (Concept)".

Feminine Gender (Concept)
<http://purl.org/linguistics/gold/FeminineGender>

Thing

- └ Abstract
 - └ Linguistic Property
 - └ Morphosyntactic Property
 - └ Gender Property
 - └ Feminine Gender

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 [submit a usage note](#)

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 ArabicNumeralGender or RomanNumeralGender 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 [submit an example](#)

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 Language Code: [chh](#)

At the bottom of the browser window, a search bar contains the text "Find: Mixer" and navigation buttons for "Next", "Previous", "Highlight all", and "Match case".

The screenshot shows the ISOCAT web interface in a browser window. The browser tabs include 'iGoogle', 'ISOCat - Data Category Registry', and 'ISOCat - Web interface'. The address bar shows the URL 'http://www.isocat.org/interface/index.html'. The interface features a sidebar with 'My Workspace' containing various project folders like 'Public', 'Thematic Views', 'Athens Core', etc. The main content area displays a table of audio resources with the following data:

#	Name	Version	Administration stat	Registration status	Check	Type	Owned by	Scope
3597	speaker ID	1:0	private	private	✓	open	Wright, Sue Ellen	public

Below the table, a detailed view of the 'FeminineGender - 1:0' concept is shown. It includes sections for '2.2 English Language Section', '2.2.1 Name Section', '2.2.2 Definition Section', and '2.2.3 Note Section'. The definition states: '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]'. The note mentions its relationship to the General Ontology for Linguistic Description (GOLD).