Preparing your Corpus for Archival Storage, LSA 2016 Workshop

We are grateful for the funding supplied by the National Science Foundation (BCS #1549994) which will make this special session of LSA 2016 possible. The session will take place on Thursday, January 7, 2016, Starting at 8am, before the start of the Annual LSA Meeting in Washington, DC.

Organizers

Malcah Yaeger-Dror, University of Arizona
Christopher Cieri, LDC

A note from the organizers

As you all know, NSF has emphasized that we should prepare our corpora for storage, so that other researchers [or we ourselves, at a later date] can use the older material for comparison with newer studies. This meeting will present critical factors that could not be included in an earlier NSF-supported workshop which examined other factors which must be considered when preparing data for comparison and sharing.
Below is the line-up of scheduled presenters. There will be six oral presentations plus discussion among the presenters and workshop participants. Since NSF is funding the workshop, there will be no additional registration fees for those already taking part in the annual meeting.

CONGRATULATIONS!
****

Students who are about to carry out their own fieldwork, or who have begun doing so, were eligible to apply for funding to help defray the extra costs of arriving early.

Student Scholars

Congratulations to the 10 recipients of the NSF sponsored scholarships to attend this satellite workshop being held at the LSA.
Joseph Kern-- University of Arizona
Jason Schroephfer -- University of Texas, Austin
Kimberly Johnson -- University of Texas, Arlington
Sarah Lundquist -- University of Wisconsin, Maddison
Daniela Benner -- Rice University
Jenny Lee -- Harvard University
Michael Lanozzi -- University of Western Ontario
Lyskawa-- UMD/ U Toronto
Kohlberger -- University of Leiden
Lily Schaffer-- U Colorado

Abstract

An NSF supported satellite workshop on Archival Preparation will be held in conjunction with, and immediately preceding, the annual meeting of the Linguistic Society of America on January 7, 2016. NSF policy now stipulates that investigators are expected to make available to other researchers the primary data created or gathered under NSF grants. However, the metadata presently associated with archived data are often inadequate to permit data (e.g., sound files) from related studies to be compared; without an agreed-upon coding protocol, there can be no effective sharing and comparison across speech corpora. Invited speakers will discuss specific coding conventions for such factors as socioeconomic and educational speaker demographics, language choice, stance and footing. Choosing appropriate metadata for these factors will facilitate sharing of corpora and research to determine how each factor impacts on language use.
NSF previously supported a workshop (at LSA 2012) in which leading scholars discussed data protocols, obtaining ethics board approval for human subject research, and ensuring that the information gathered about human subject demographics, attitudes and the situations in which they were recorded provide enough scope and detail to permit meaningful comparison across studies and thus encourage data sharing. Following that model, this workshop will extend the topics covered and provide a training forum in which to develop protocols for sharable data that conform to the spirit of NSF policy. This award will also support the participation of students in the training and discussions.

Outline

The workshop will be composed of four sections.

1. The first section will propose appropriate coding options for socioeconomic status and education.

The two speakers will be:
Anne Fabricius, Roskilde University
Social class, social capital, social practice and language in British sociolinguistics: unraveling historical and ethnographic complexities
This discussion will elaborate on the approaches to social class analysis and coding that I and others have pursued in studies of the elite/establishment sociolect of England over the past fifteen years. Changing social and political contexts as well as 'class as an ideological construct' within British society have ramifications for sociolinguistics and corpus work. We will look at several traditions of social class analysis and their potential contributions to sociolinguistic research. The importance of fine-grained historical and contextual understanding will be a recurring theme.

Suzanne Evans Wagner, Michigan State University and Robin Dodsworth, NCSU
Conceptualizing and coding social class in North American sociolinguistics
North American sociolinguists have classified speakers by their (perceived) social status in three main ways: (i) indices of occupation, education etc; (ii) locally relevant categories like 'Burnout'; (iii) evaluation of the 'linguistic market' value of speech. These methods will be evaluated in conjunction with network analysis, with a view to establishing future coding systems that allow for both geographic and longitudinal comparisons across datasets. To illustrate this discussion, examples are drawn from recent studies of three US metro areas and from archival speech data.

2. The second section will propose coding options for specific situational variables.
Richard Ogden, York University, UK
Coding categories relevant for interaction
This contribution looks at how interaction shapes (and is shaped by) language. Starting with a brief overview of how interactionally relevant categories have already been coded, we look at the implementation and organization of social actions through talk, focusing on the use of clicks in spoken English conversations. Clicks are common in displays of affect (both positive and negative); we will consider how such displays are made relevant and implemented in conversation, and how relevant details of interaction can be coded, including matters of sequential organization and action type.

Frans Gregersen, LANCHART Centre of Copenhagen University
Discourse contexts within sociolinguistic interviews, a presentation of the LANCHART DCA coding scheme
It is the hallmark of a mature science that previously collected data and results are tested against new knowledge. In this endeavor, a lot depends on the quality of metadata. Much of this information is commonplace and refers to technical details about how data were collected, recorded and transcribed etc. But the intelligent exploitation of older data will crucially also depend on how to document variation within recordings. In my presentation I will review what we have done at the Copenhagen University LANCHART Centre to control for internal variation within the recording sessions and critically discuss the resulting coding scheme.

3. The third section will propose coding options for bilingual /code switching segments.
Naomi Nagy, University of Toronto & Paulina Lyskawa, U Maryland
Moving forward with multilingual transcription
Since 2009, the Heritage Language Variation and Change in Toronto Project has been building corpora of conversational speech in a range of Heritage Languages in Toronto. Teams of students from each community have developed language-specific transcription protocols. We apply variationist methods to quantify the effects of various contextual forces on the selection of forms both within and across languages. This presentation will describe and problematize how we indicate use of multiple languages within one conversation and efforts to maintain consistency across protocols from different languages/communities, commenting on efforts to make these transcripts useful for inquiries developed subsequent to transcription.

Barbara E. Bullock & Almeida Jacqueline Toribio, University of Texas, Austin
Toward automated methods of bilingual annotation
Coding of bilingual data requires levels of linguistic annotation and metadata that are typically irrelevant for populations in which the speakers sampled are presumed to be dominant and proficient in only one and the same language. As such, linguists with interests in bilingualism have traditionally had to manually categorize the language of the data, the type of data (code-switching, borrowing, calquing), and the linguistic abilities of its speakers. Here, we discuss the procedures and the results of the Bilingual Annotation TaskS (BATS) Force to automatically classify bilingual language data in ways that can be scaled up for large data sets.

We have been lucky enough to have also 'snared' Jacob Eisenstein from Georgia Tech, who will both deal with social media metadata, and add a new perspective on the issues being discussed in the other sessions:

Jacob Eisenstein, Georgia Institute of Technology
Social Media Metadata as Sociolinguistic Evidence
==========
The increasing ubiquity of social media offers exciting new opportunities for sociolinguistic research, due to three major advantages: (1) the unprecedented reach of these resources, which enable aggregation across millions of individuals; (2) the possibility of studying language in use in a multitude of situations with real social stakes; and (3) the unprecedented availability of metadata, which adds crucial context about the individuals under study and the social situations in which they are engaged. Specifically, this talk will address social media metadata relating to time, geography, and social networks, describing the linguistic research that this metadata enables: tracking individual and community language evolution; linking written and spoken variation, in both their linguistic and socioeconomic dimensions; and identifying large-scale evidence of style-shifting based on social network factors. I will discuss specific social media sources (mainly Twitter and Reddit), and will also discuss new research challenges raised by these data sources, as well as some ongoing work on how these challenges may be addressed.

We hope that all members of the group will have exchanged papers by (at least) 12/25, and will be happily ensconced at the meeting site, and will be ready to compare notes with the other authors, as well as with the 10 students who will be officially designated Student Scholar Attendees for the event, over dinner on Wed night. If possible, we would like preliminary versions of papers to be available for us to post on the workshop website. As you see, all speakers will be given 40 minutes to make their points. If you have a software program that is relevant and that is sharable, we will post that as well. There will be ample time for discussion both during and after the workshop..
All of the invitees [including the ten students] are cordially invited to a working dinner, the night before the workshop [Wed, Jan. 6].
-
Table 1. Plan for a 2016 workshop, Starting Thursday Morning 1/7.

Session	Time	facilitator	presenters	Title	Creating sharable corpus & archive
1/6	7:30		Working	Dinner	Place to be determined.
Coffee	Break				
1	8:30-	Cieri			Session I – Socioeconomic Coding
	8:30		Fabricius		The British View
	9:10		Wagner/Dodsworth		The US View
	9:50	Cieri			Discussion & Resolution
Break	10:00				
2	10:00	CMC/myd			Session II – Situational Variation
	10:00		Eisenstein	Metadata as Linguistic evidence	Media metadata as sociolx evidence
	10:40		Ogden	Conversational Discourse	The CA codes for interaction
	11:20		Gregeresen		The Lanchart codes for IVs
	12:00	CMC/myd			Discussion & Resolution
Lunch	12:15		Working	Lunch	With coffee at the end....
3	1:15	Yaeger			Session III--Code switching
	1:15		Bullock/Toribio	Metadata for bilingual coding	Bilingual coding
	1:55		Nagy/ Lyskawa	Metadata for bilingual coding	Bilingual coding

Break	2:30				
	3:00-	Cieri	all		Concluding Discussion

--
The authors of papers in a given section will be sharing their papers with each other, and we will try to get all of the ppts to all of you AT LEAST BY DEC 25.

WE ARE LOOKING FORWARD TO THIS SESSION, AND HOPE YOU ARE AS WELL!