Linguistic Data Consortium

An evolving perspective on data coding conventions

Malcah Yaeger-Dror,

malcah@email.arizona.edu

Sociophonetics

Linguistic Data Consortium

Coding Conventions for Data Collection for Archival Sharing

Demographics

Situations

Attitudes

--A users' view

Filling our cells…

- 1. However many groups we have they should all be represented
- 2. e.g, If we have 4 women in a group,
  - we should try to have 4 men.
- 3. If we have a clue that a group might speak distintively,
  - We can only prove it by coding for it, and filling the relevant cells.
- In this presentation I want to suggest ways in which we can make our coding conventions more specific, to
  - facilitate a researcher’s choice of data
  - Permit a focus on inclusion of *appropriate* speakers in a sample.

Linguistic Data Consortium

let's define 'METADATA'

- **1. Structural metadata- (how an archive is to be built)**
  - the design and specification of **data structures,**
- **2. Descriptive metadata-**
  - individual instances of data - **metacontent.**
- **DIGITAL data using standards specific to a discipline**
  - **This increases the usability/sharability of the data**
- **In this presentation I want to suggest ways in which we can make our coding conventions more specific, to**
  - **facilitate a researcher's choice of data**
  - **Permit a focus on inclusion of *appropriate* speakers in a sample.**

Today's Metadata Foci

- I have been formulating coding conventions for speech archives. There are three foci for coding on which I will try to elaborate.
  - **1. LDC's demographic coding** has been upgraded constantly to reflect linguists' needs. As a sociolinguistic user of LDC corpora I have a few additions to the coding to suggest, so they could be used as relevant research criteria in future studies.
  - 2. As a sociolinguistic user of LDC corpora I have found that most of the **situational information** is well spelled out for any given corpus, so with a few exceptions, coding for the social situation could be almost automatically inserted into the record for each corpus.
  - **3.** I will also discuss evidence from recent studies which are demonstrating the influence of **interpersonal attitudes on speech variation**, and most of the talk will focus on the speakers' attitudes toward their interlocutors, and how we might be able to go about determining this information honestly without recourse to Gilesian psychological studies.

LDC Linguistic Data Consortium

DEMOGRAPHIC Coding Conventions

Demographics

Situations

Attitudes

--One user’s view

Demographics 2B coded

- **(Dialect) Region**
- **Sex** (M/F) – or sexuality {mm/mf/ff/fm}
- **Birth Date**
- **Age** @ IV
- **Race**
- **Heritage** group {how far back? How mixed? How strongly identified?}
- **Religious** affiliation
- **Socioeconomic** background

Demographics

- **Age** (@IV--α place within the culture)*
  - Til age 5
  - Ages 5-12

  ***Can later be used to distinguish CHANGE from AGE GRADING**

  - Teen years: Youth Culture (inferable in indigo)
  - Wage Earning years [when the ML is in effect]
  - Post retirement [when ML may no longer be in effect]

  *cf. the recent post on *LgLog*: Mel Brooks claims that, based on observations of his Brooklyn neighbors, he believes that speaking Yiddish is something that happens to you when you get old. KevinM hopes that his daughter will lose the '*Like…*'quotative

Demographics *not* always included

And not always inferrable:

Region(s) where speaker has lived

- Til age 5
- Ages 5-12
- Teen years– ‘youth culture’
- Wage earning years
- Retirement years

And the speaker’s attitude toward that region.

Linguistic Data Consortium

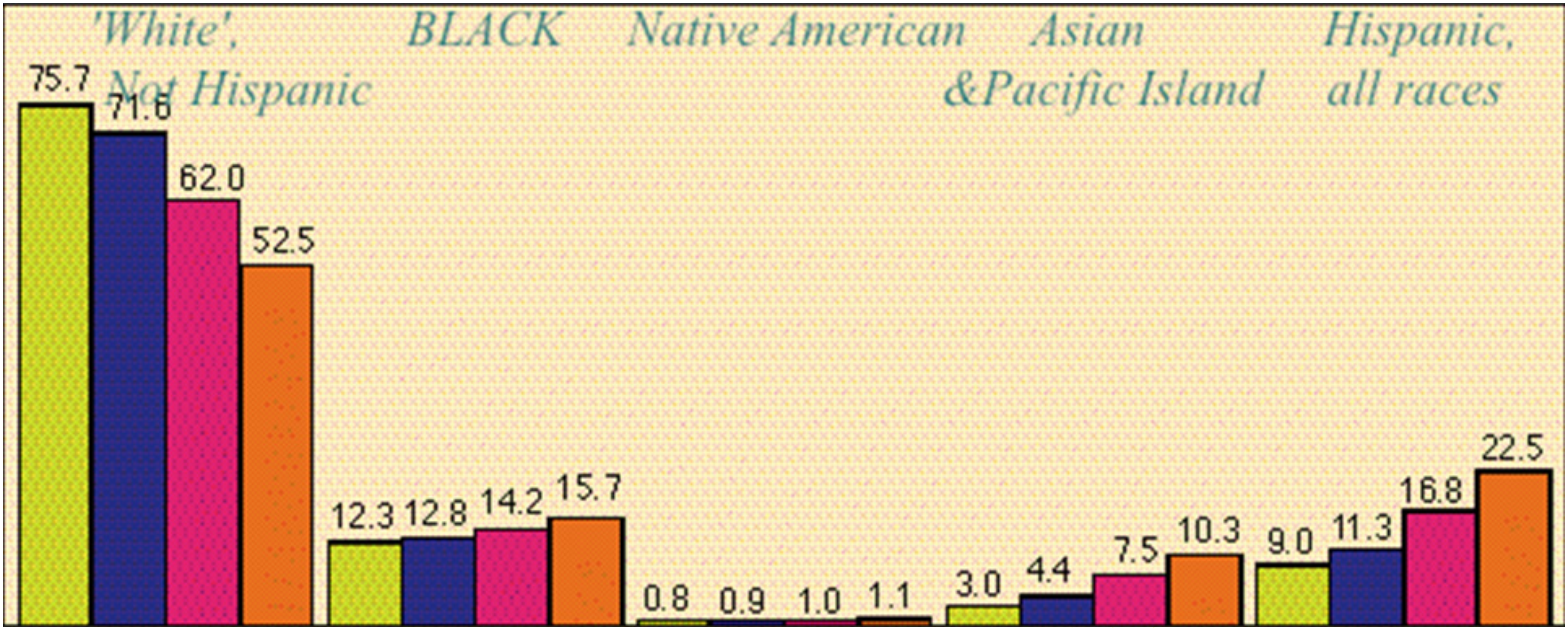
Demographics *not* always included

And not always inferrable:

- ‘**Race**’/Ethnicity {African, Chinese, Japanese, Desi…}
- **Family origin** Ethnicity {Italian, Hispanic….}
- **Religion**/ ‘Ethnicity’
  - {Muslim, Copt, Druze, Catholic, Pentecostal, Amish....}

US Census Projections

Percent of Population 1990, 2000, 2025, 2050



http://www.census.gov/population/www/pop-profile/natproj.html

‘Philadelphia’s changing makeup’

? Is there a ‘tipping point’ for community cohesion?

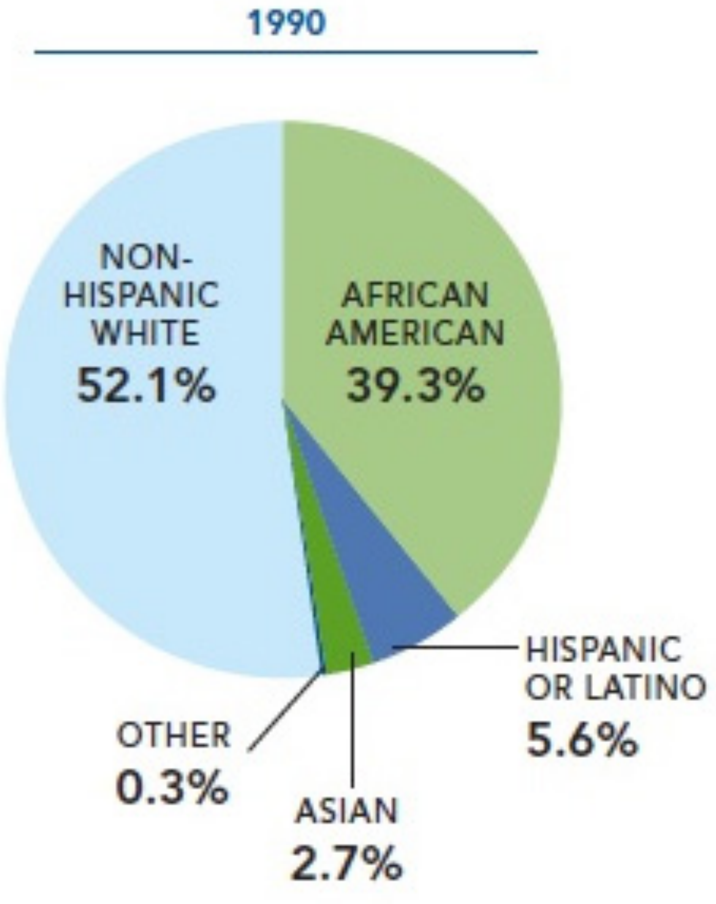
Over the last 20 years of Philadelphia demographics

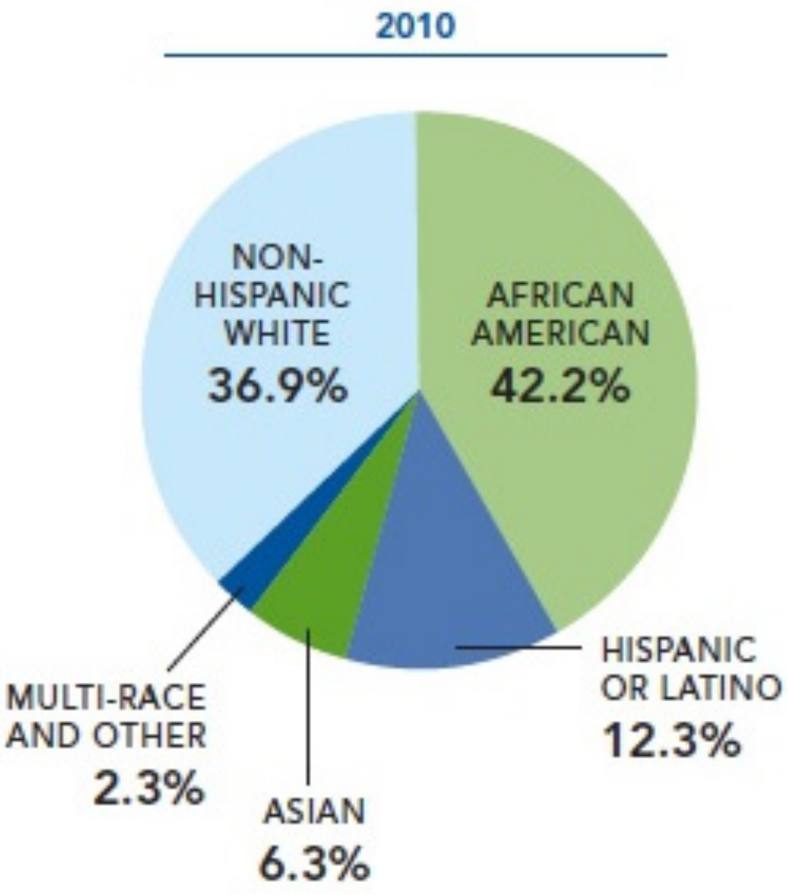
- ‘**Whites**’ –ethnic or other- 31.9% *decrease*
- **‘AAE’+ others—3% increase** [14% *decrease* in University City]
  - {Old Phily, new Philadelphian, Caribbean, Dominican, Sudanese, etc…}
- **Hispanic- 110% increase** [84% in University City area]
  - {major split: Mexican vs. Puerto Rican, +}
- ‘**Asian**’ **-127% increase** [104% in University City area]
  - {Desi, Hmong, Cambodian, Thai, all Chinese, Japanese, Shanghai}

--[*DP* front page, 6/9/11]

Linguistic Data Consortium

Demographic Change: 20 Years in Philadelphia





Linguistic Data Consortium

Population 'Balance'

- There is a consensus that exemplar dynamics
- [who you talk to / or hear {in person, on TV, on the bus…}]
- Has a strong influence on dialect usage
- Both Pew and Mumford Center have demographic information like that on the web for different neighborhoods of different cities…
- So, we should always specify the population balance of
  - Schools
  - Work places
  - City
  - Friendship network

What does that mean
to an average speaker-hearer?

Note that if we think of it not as numbers but as what percentage of the folks you talk to have specific speech characteristics, we can say that now

- Almost half the people you interact with in Phily are ‘Black’,
  - although that doesn’t tell you how many have AAE characteristics.
  - & in certain neighborhoods that entails that most speakers will be ‘Black’
- You’re more than two times as likely to talk to ‘Asians’ or ‘Latinos’
  - Although the percentages are not yet likely to have an ‘exemplar based’ infl.
  - And you may not even be registering their speech as ethnically marked.
  - And (again) there’s no evidence for how marked their speech is,
  - Or even if all speakers have similar characteristics.– *e.g.*,
  - PR ‘Latinos’ may no longer have strong L2 characteristics, while
  - Mexicans may have quite different *ad stratal* features

Linguistic Data Consortium

Demographic Drawbacks

- The Pew and Mumford statistics have a drawback:
- Many sociophonetic studies have similar drawbacks.
- Are all ***'Black'*** Speakers African American ?
  - {Caribbean, Dominican, Nigerian, Sudanese, 'Ethiopian', South African…}
  - {Blake, Shousterman…Rickford}
- Are all ***Latino*** speakers from a uniform group?
  - {Chicano≠Puerto Rican≠Mexican≠Dominican…}
  - {Fought, Zentella, Mendoza-Denton}
- Are all ***'Asian'*** speakers from a uniform group?
  - {Desi, Bangladeshi, Untouchables,…}
  - {Japanese, Chinese, Korean, Taiwanese, Tibetan, Hong Kong, Vietnamese}
  - {Hall-Lew, Wong}

Demographic Drawbacks

- For that matter, are all 'whites' homogenized after one generation?
- The issue of ethnic heritage [and how long it takes to be neutralized]
- E.g, In my hometown, there are Catholic parochial schools
- Until a few years ago {Polish, Irish, Italian}-heritage students went to separate schools
- In South Philadelphia {Irish, Italian}-Heritage students share a school
- And yet, like with Eckert's and Mendoza-Denton's teens, ethnicity of their grandparent/great grandparent generation was still significant.
- And marked by sartorial distinctions {hair, ties, belts…}
- Do they talk alike? We don't know.
- We won't find out unless that ethnic heritage distinction is coded for.

Religious Demographic

- In South Philadelphia {Irish, Italian}-Heritage students share a school
- How much more different are students of similar heritage who do not share the same social networks & schools in their youth?
  - In Tucson there are two Chicano-Hispanic groups who do not mix
  - {Catholic/Pentecostal}
- Returning to the 'Asian' speakers, which is more important:
- Great- grandparent L1 or Great – grandparent ***religion***?
  - {Muslim, Hindu, Parsi, Buddhist, Shinto, Confucian, Baptist}
  - {Wahabi, Druze, Shi'a, Sunni, Sufi, Copt, Maronite, ... }

? Can we say that the more likely it is that identity maintenance focus is on religion, the more critical it is to code for religion?

- And (how) do we isolate religion from {SES/ML/...}

Class(ic) Alphabet Soup

How do we code for what ‘callers’/ ‘speakers’ DO for a living?

- Years of Education [found most useful in Tehran studies]
- ‘Class’
  - **Type of employment [blue/white collar]**
  - SES
  - ML – Linguistic Marketplace (correlates with social network as well)
  - **> Perhaps**

? most advisable to say what speakers do for a living, and leave the interpretation of that information to those who know the community themselves, and can interpret it more appropriately?

? Or with notes specifying what the hierarchy is in that community?

Linguistic Data Consortium

Demographic recap

- **Region** (often just one!)
- **Sex** (M/F) – or sexuality (mm/mf/ff/fm)
- **Age (best coded by b/d** - reveals change in apparent time)
  - Just the actual b/d goes into the file
  - Later work can distinguish what the cultural cut off is for rapid change.
- **Age-grading** within the culture (age@IV)
- **Ages and settlement patterns** (*cf.* Feagin)
- **Race**
- **Heritage** group {how far back? How mixed? How strongly identified?}
- **Religious** affiliation
- **Socioeconomic** background

Linguistic Data Consortium

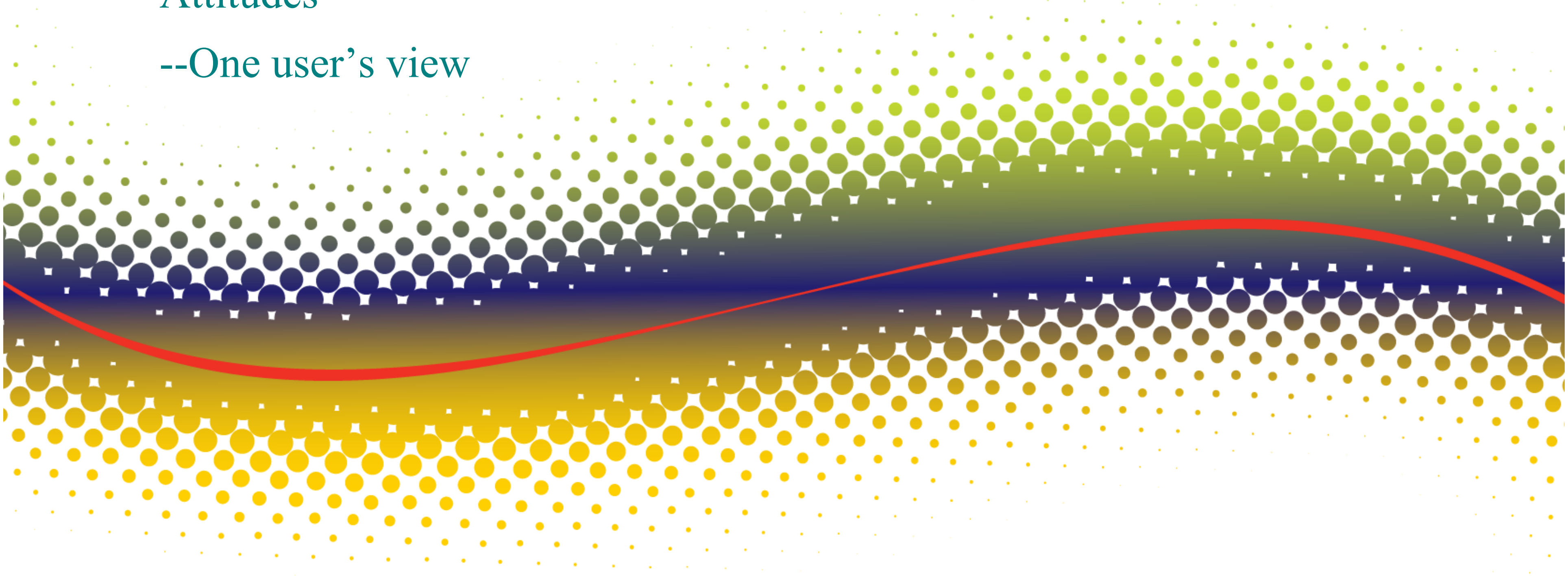
SITUATIONAL Coding Conventions

Demographics

Situation

Attitudes

--One user's view



Social Situation

Demographics? Or social situation?

Interaction between/among participants

- **Relative Gender** (4-way):
  - same or different?
  - and does it matter here??
- **Relative Age:** who is older, and does it matter here?
- **Relative 'Class' or education or earning power**
- Relative Family origin Ethnicity {*e.g,* Italian, Hispanic….}
- Relative Religion 'Ethnicity' {*e.g,* Muslim.., Amish….}
  - (*cf*. Schegloff on formulating place)
  - (*cf*. Giles' perspective on intergroup relations)
- **Culture(s) of the interlocutors**
  - (*cf*. Hofstede's **Framework for Assessing Culture**
  - (*cf*. http://en.wikipedia.org/wiki/Geert_Hofstede)

Demographics? social situation?

Interaction between participants

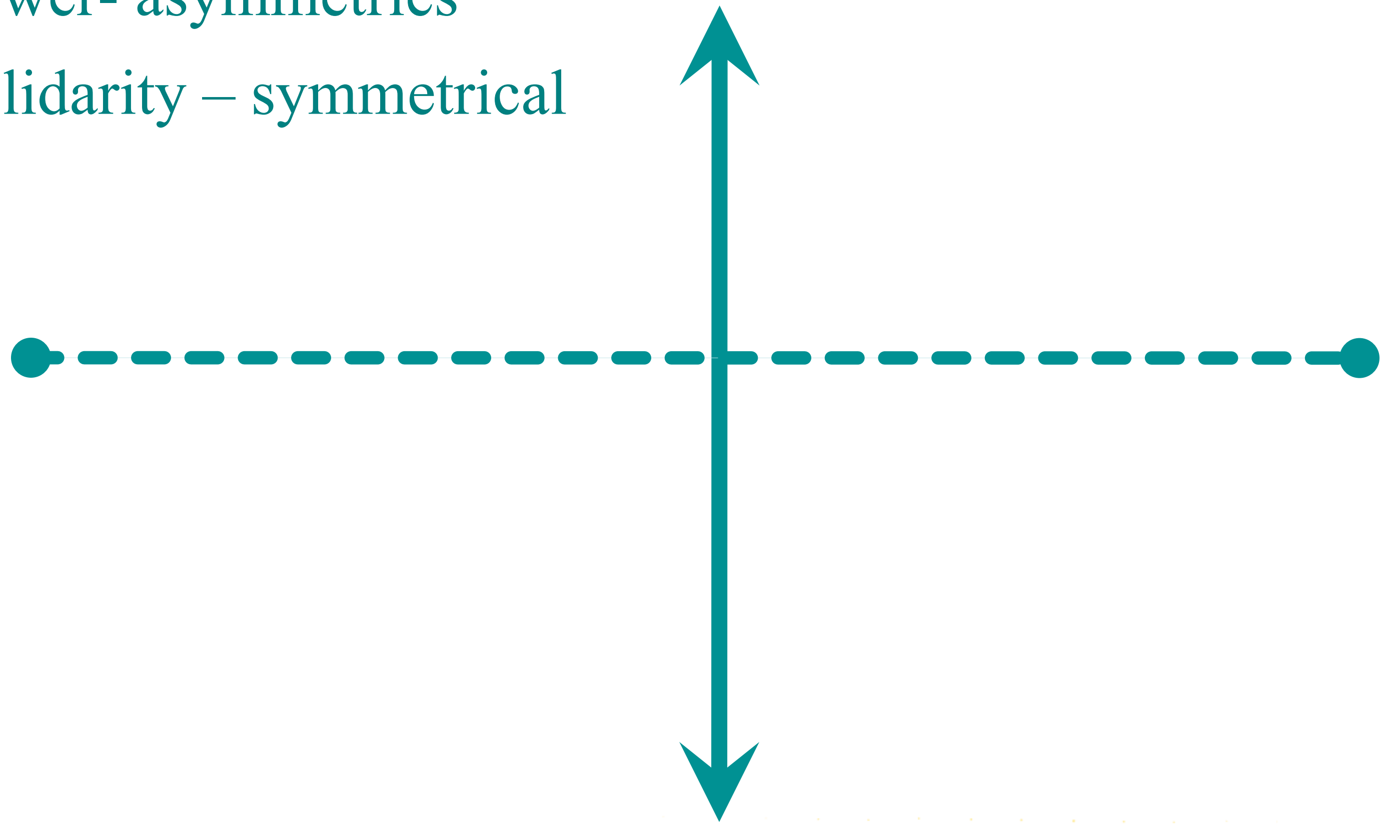
Issues of relative POWER and SOLIDARITY

(if they are sufficiently acquainted)

-Brown & Gillman 1960

Power & Solidarity

- Power- asymmetries
- Solidarity – symmetrical



- Both vary within and across cultures as well as within the dynamics of an individual interaction.

Demographics? social situation?

Interaction between/among participants

These are locally/culturally variable --

- **Relative Gender** (4-way):

  same or different,

  and does it matter here? To whom does it matter?
- **Relative Age:** who is older, and does it matter here? To whom?
- **Relative 'Class' or education or earning power:…**
- **Relative Race:….**
- **Relative Family Origin/** Ethnicity {Italian, Hispanic….}:…
- **Relative Religion**/{*e.g,* Muslim, Jewish, Catholic, Amish….}:…
- **Relative** Ethnicity/ies of the interlocutors
  - (cf. Schegloff on formulating place)
  - (cf. Giles' perspective on intergroup relations)

Hofstede: Cultural Perspective

- **Culture(s) of the interlocutors**
- **Low~High Power /Distance cultures**
- **Uncertainty avoidance cultures**
- **Masculine** [power] ***vs*. Feminine** [solidarity] **cultures**
- **Individualistic *vs*. Group-focus cultures**
- **Uncertainty avoidance cultures**
- **Quantity *vs*. quality of life cultures**
- **'Time Horizon':**
- **Long term *vs*. short term focus** (so-called).
  - **Long term α power, shame,**
  - **Short term α face,**
  - (*cf*. Hofstede's **Framework for Assessing Culture)**

Hofstede Table

Country	Power/ distance	Individualism	Masculinity	uncertainty avoidance	long term perspective
Japan	54	46	95	92	80
South Africa	49	65	63	49	
US	40	91	62	46	29
Australia	36	90	61	51	31
U K	35	89	66	35	25
Ireland	28	70	68	35	
Jamaica	45	39	68	13	
Venezuela	**81**	**12**	**73**	**76**	
Mexico	**81**	**30**	**69**	**82**	
Colombia	**67**	**13**	**64**	**80**	
Ecuador	**78**	**8**	**63**	**67**	

Judgement of sample size

- **What do we code for?**
- **What do we stipulate, to insure limiting of sample size?**
- **Do we want to include all these factors as ‘variables’**
  - **(Each new factor doubles the size of the corpus needed.)**
- **OR: Do we want to limit corpus size**
  - **(avoiding factors peripheral to a specific set of interests?)**
- Even analysis of one sociophonetic variable is very time consuming
- One may prefer to compare variation in a number of variables,
  - to see if they covary
  - or if there is a pattern which distinguishes different features.
- **This is not a plea to code for everything**
- **But to pick one’s “battles” judiciously,**
- **And verify that other features are recoverable.**

Judgement of sample size

- The situational variables discussed here are usually ignored.
- Ignoring them is safe if they are held steady.
- But they should still be stipulated somewhere.

OLAC Genre/Register Codes

olac term	# parties				
drama=	2+ party f2f?				
formulaic discourse=	single or grp	prayers	curse,blessing	fables/stories	formulae
interactive=	2+ parties				
lg play=		jokes	secret lg	coded	riddles, etc
oratory=	single party	speeches	lectures	invocations	semons
narrative=	single party	story telling			
report=	single party	news,	class	journal	dry run etc
singing=	single or grp	chant	**song genres**	chorus	opera operetta
unintelligible=	single or grp	glossolalia			

Hymesian genre

1	2	overhearer	participants	relations	among	
singular	singular	singular	speech {}		rel power	solid
plural			f2f			
institutional			power?			
unidentified						
acquainted				interactive		
		Intended?				
education	…					
profession	…					
role	…					

Hymesian register/genre

channel	production	setting	purpose(s)	stance	footing	topic
speech	scripted	shared±	narrate/report	none	none?	immediate
writing	edited	private/ public	inform/explain	neutral	neutral?	global domain
signing	iff cyff	specifics?	describe	Supportive /	adversarial	status of person discussed
recorded {radio,tv..}		place	persuade	adversarial	supportive	religion
scripted {}		time	entertain			sports
written			educate			art
			fact			education
			{info,			etc
			{opinion			

Biber's Table for English written genres

Participants	overhearer
	addressor
	addressee
relationships	interactive
	roles
	relations
	Shared knowledge
channel	written
	oral
production	
comprehension	
setting	time
	Place…

Biber's Table for English oral genres

- 3 Dimensions are on a contiuum (based on specific syntactic or lexical features in the text)
- Dimension 1 is an oral/literate continuum [90.7%]
- Dimension 2 is a procedural/content continuum [51.7%]
- Dimension 3 is ‘reconstructed account’ [story telling?] [20.8%]
- Dimension 4 is ‘teacher centered stance’. [45.2%]

Social Situation…

- When it comes to *genre* or style Most of the OLAC or Hymes situational variables can be coded for older corpora, because each corpus is composed of ONE 'style', or with a short intervention by another:
- This is both the strength & weakness of sociolinguistic corpora.

Interpersonal-interactive variables [as described above]

- Based on the interaction between speaker demographics, and
- Based on each speaker's cultural evaluation of
  - interlocutor demographics
  - As vs. his/her own

are only rarely included (in recent studies)

And cannot generally be determined later.

Linguistic Data Consortium

SPEAKER ATTITUDE

ATTITUDE Coding Problems

Demographics

Situations

Attitudes

--One user's view

Linguistic Data Consortium

Exemplar dynamics vs. Attitudinal studies

- In recent years there has been increasing evidence that speakers' social attitudes – which are partly based on the interaction of individual speaker demographics, cultural patterns, and the type of situation which is being recorded also influence speech, sometimes quite radically, even influencing language choice (much less finer dialect variation).
- However, following older data-gathering protocols, most research groups (that LDC relies on for our own protocols) have not yet begun to study these variables in a systematic way.
- In tandem with this limitation on present sociolinguistic methodology, Experimental/Laboratory Phonology has recently been dominated by the theory of exemplar dynamics.

Sociolinguistic theory vs Exemplar dynamics

- Older sociolinguistic theory maintained, with developmental psychologists, that your dialect was ‘frozen’ by the time you were 12ish. This would obviate not only the need for a study of attitudes, but even some of the ‘demographic’ information proposed above.
- Pierrehumbert (2001/2/3) has postulated that a speaker’s phonology varies relative to the number of tokens of a variable s/he hears with a given realization. This is also a recent ‘default’ position for lg. change voiced by Trudgill (2008) among others, and is consistent with Milroy’s (1980) theory of speech ‘networks’.

Linguistic Data Consortium

Exemplar dynamics

This theory neatly conflates

- who you talk to [social network], &
- who you hear from the environment,

with change past adolescence easily accounted for as

- Older speakers become more scarce
- Younger speakers dominate the airwaves & neighborhood
- Speakers change their social groups/networks

One recent study (Kammacher *et al* 2011) refers to this as the 'napoleonic theory' of language variation: or the **Napoleon Principle** (Brink & Lund, 1979:202).

"C of P"

- **Community of Practice** studies are not necessarily embroiled in this conflict, since [merely] changing your social network may alter your linguistic choices without the intervention of 'attitudes'.
- This theory of Community of Practice, however, does involve determination of the degree of '**authenticity**' of a speaker: that being the degree to which the variation can be found to be '**below the level of conscious awareness**'.
- So while there is an increasing number of studies finding change in real time [past adolescence], the motivator ('motor') of phonological [and morphosyntactic/lexical…] change is still under-determinable and not determinable from the data available.

Attitudes count

- An increasing number of studies are finding, that the perspective of Howard Giles should not be ignored, because speaker attitudes influence the way they talk.
- And how individual speakers [and whole communities] alter the way they speak over time.
- As we learn more, it seems that both analysis and recognition studies will benefit appreciably from a clearer understanding of the as-yet-uncoded attitude information.

Attitudes count

We will ignore many recent studies of dialect variation (even those which are studying variation past adolescence) which have been agnostic on the critical issue of whether

- the change is Napoleonic/Exemplar/Trudgillian, or
- Speakers are influenced by their attitudes toward a feature that they hear to either accept it, or not.
- The rest of the talk should focus on different studies that are analyzing for social attitudes.

Ethnic Orientation

- Parents’ country of origin
- Length of residence [locally]
- Age at arrival
- Age began learning local lg.
- Where was the Local Language [LL] learned?
- Preferred lg ? [for which domains?]
- Lg choice – both of speaker with X, and of X with others
  - With parents
  - With sibs
  - With friends
  - With significant other.

Linguistic Data Consortium

Ethnic Orientation

- Other questions found significant are going to be harder to get past the IRB, perhaps, although they've been found to be significant:
- Could you go back to your 'homeland'?
- Would you go back to your 'homeland'?
- What likelihood is there of going back there?
- What political party do you vote for?
  - (Republican Hispanics try harder in the US, not measured elsewhere)
- Congruency between religious persuasion, & politics, ethnicity, attitudes?
  - *E.g.*, Chinese Baptists the same as Confucians?
  - *E.g.*, Hispanic Pentecostals the same as Catholics?
  - *E.g.*, Copts the same as other Cairenes?
  -

Exemplar Dynamics vs. Gilesian studies.

Coding Conventions

- Coding Conventions for Archival Sharing.
- While at LDC I have been formulating coding conventions for speech archives. There are three foci for coding on which I will try to elaborate.
  - 1. As a sociolinguistic user of LDC corpora I have found that most of the **situational information is well spelled out for any given corpus, so with a few exceptions, coding for the social situation could be almost automatically inserted into the record for each corpus.**
  - **2. LDC's demographic coding has been upgraded constantly to reflect linguists' needs. As a sociolinguistic user of LDC corpora I have a few additions to the coding to suggest, so they could be used as relevant research criteria in future studies.**
  - **3. I will also discuss evidence from recent studies which are demonstrating the influence of interpersonal attitudes on speech variation, and most of the talk will focus on the speakers' attitudes toward their interlocutors, and how we might be able to go about determining this information honestly without recourse to Gilesian psychological studies.**

Thank You