Challenges and Alternative Data Sets

Christopher Cieri and Malcah Yaeger-Dror

ccieri AT ldc.upenn.edu, malcah AT gmail.com

This work was supported in part by NSF Grant BCS #1144480 with supplemental funding from LDC.

Outline

- 1. Review of information about accessible corpora
  - A. Corpora within the LDC archives (CMC)
  - B. Fun corpora requiring initiative but minimal funding
    - i. Broadcast media
    - ii. Musical media
    - iii. Political media
- 2. Take away message
- 3. Conclusions

LDC

- Consortium of universities, companies, government research labs that develop, share, use language data
- Model: members receive ongoing rights to corpora published in their membership years
- To date, LDC has distributed:
  - >108,000 copies of
  - 1860 data sets (>600 published) to
  - >3500 organizations in
  - 71 countries
- Identified more than 11,000 papers the rely on LDC data
- Most data developed for human language technology R&D though much relevant for other linguistic research
- Closing perceived gap between data for HLT & Linguistics
- Most University join via Library or Engineering
- Grants in Data for worthy, impecunious students

Panel Study Challenges

- Initially, finding large, appropriate sample
  - telephone studies: 50% do nothing, 70% do 80% of requested
  - subjects refuse, flake, drop-out, ‘misbehave’
  - in a stratified sample, some cells harder to fill
- In following epochs
  - difficulty locating previous subjects
  - difficulties not necessarily in proportion to sampling parameters
  - subjects’ lives have changed other than aging
  - analytic methods have changed
  - cost for changing methods, an opportunity cost for not changing
- In analysis
  - are data representative of recording epoch
  - are different epochs truly comparable

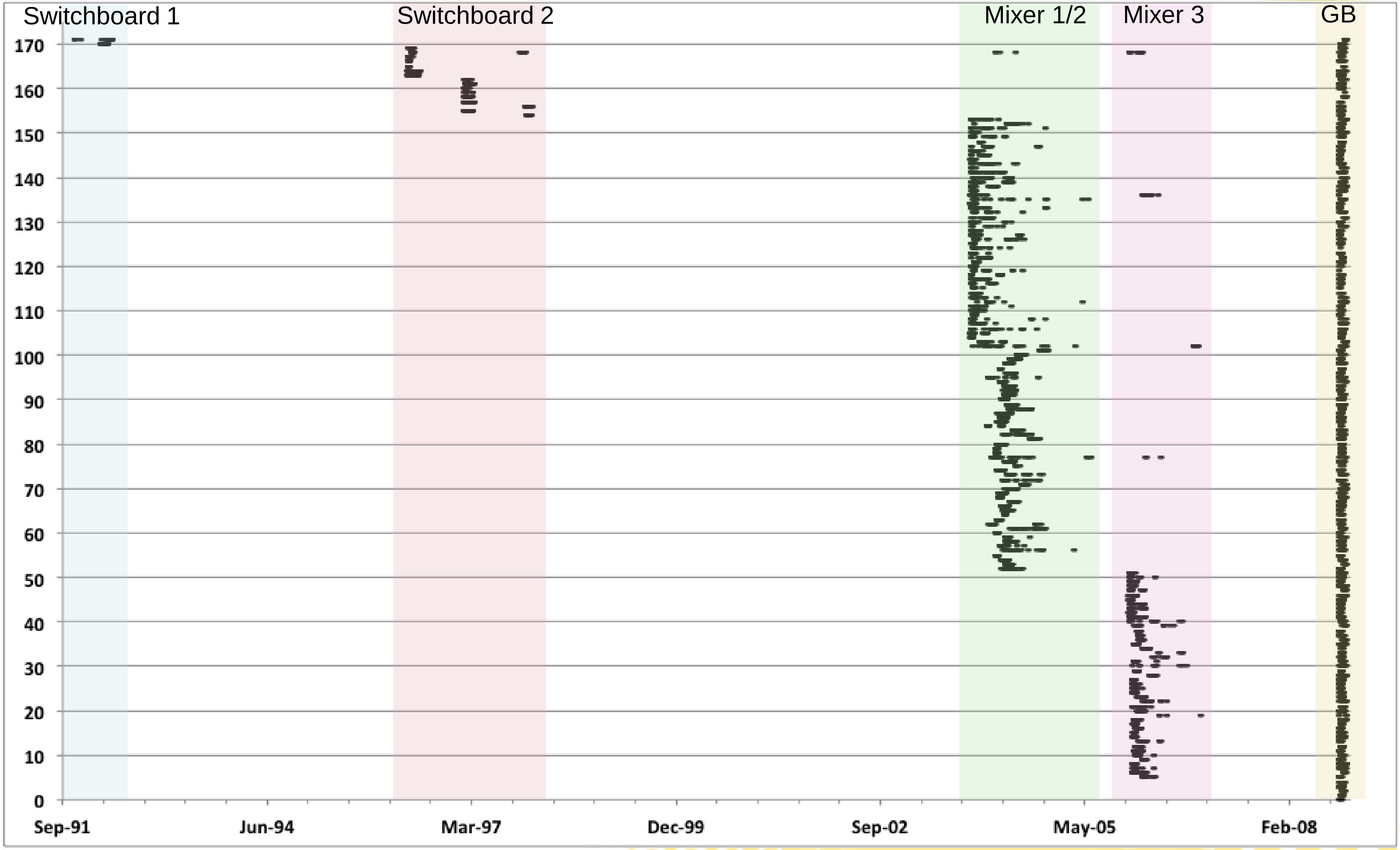
Found Data Advantages

- Reduces burden on
  - time
  - resources
- **Found Data** is more adaptable than **Found Findings**
  - can be re-annotated, re-sampled, augmented
  - if public, supports replication
  - benchmark (stable component) for competing analyses

Greybeard

- Goal: support speaker recognition technology development
  - multiple sessions per speaker
  - differing by: time, handsets, topics
- Tasks
  - find speakers who made 5+ calls in a study >=18 months earlier
  - re-contact and recruit into new study
  - record an additional 10-20 calls
- Characteristics
  - 2+ epochs
  - 8, 12, 24 calls per talker per epoch
  - 5-10 minute telephone calls, multiple handsets, locations
  - among strangers
  - topics suggested, not enforced
  - https://catalog.ldc.upenn.edu/LDC2013S05

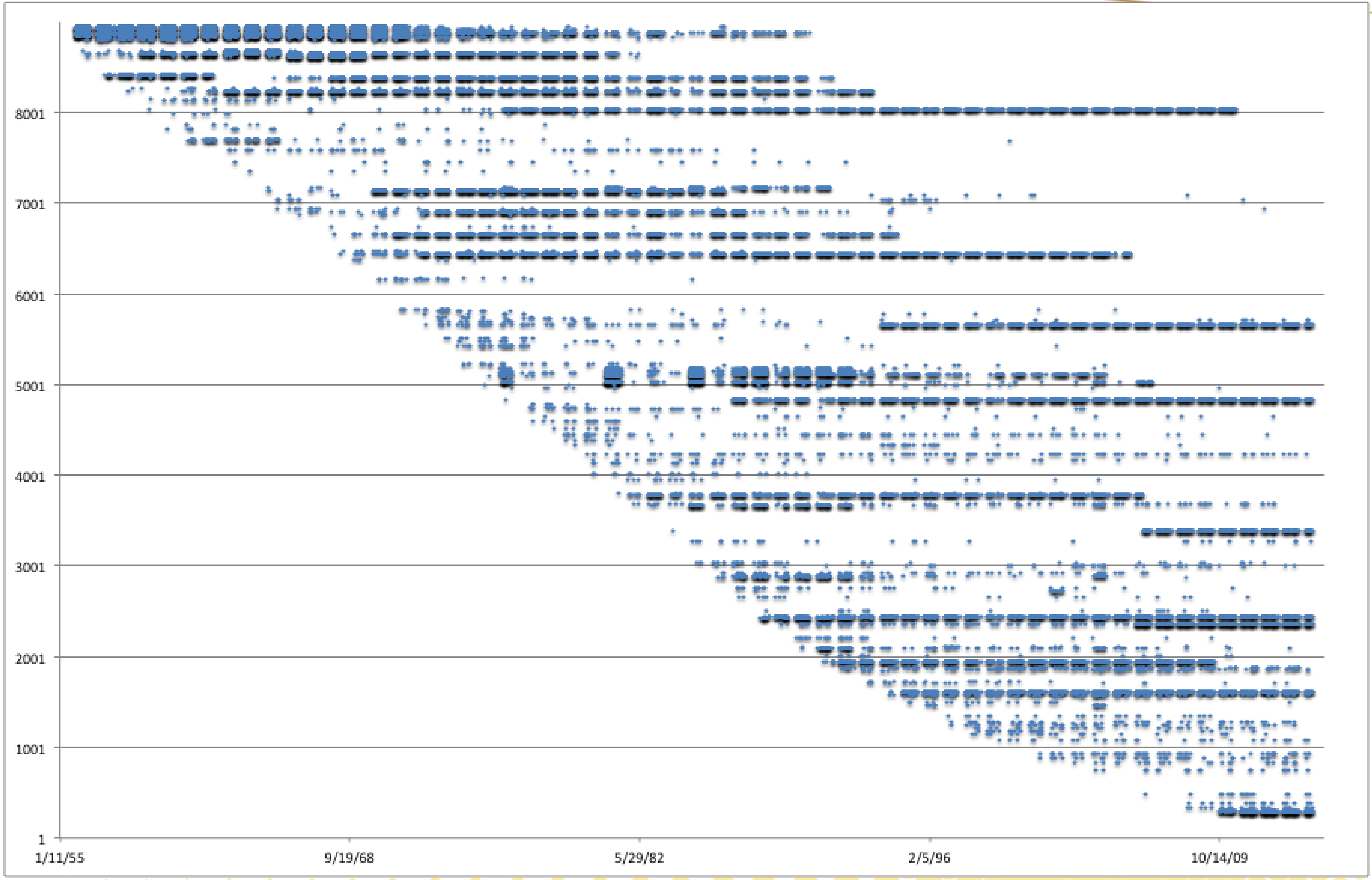
Greybeard Calls by Subject and Time



SCOTUS

- Goal: increase public access to SC oral arguments
  - www.oyez.org/cases/
  - advocates for/against cases before the SC make oral arguments
- Characteristics
  - web accessible, transcribed oral arguments
  - LDC added forced alignment and diarization (speaker / turn)
  - almost 9000 talkers
    - many sessions from a relatively small number of Justices
    - relatively few sessions from lawyers arguing a specific case
  - specific genre
- Copious demographic & attitudinal metadata on Justices
- Situation is documented
  - www.supremecourt.gov/visiting/visitorsguidetooralargument.aspx

SCOTUS by Frequent Talker, Time



Testing Comparability Assumptions: Mixer 6

- built to support robust speaker recognition
- duration: 8 months, per speaker ~1 month
- 594 speakers * 10 calls (N=4410) + 3 LDC visits (N=1425)
- all native English speakers, most Philadelphian
- phone call: reduced bandwidth but close talking
- visits: 45 minutes, 14 simultaneous microphone recordings
    - see Rathcke & Stuart-Smith, this conference, on differences in F1 by microphone
    - Repeating questions: <= 1 minute
    - Informal conversation: ~ 15 minutes
    - Transcript reading: ~15 minutes
    - Telephone call: 10 minutes
- metadata: year of birth, years of formal education, highest degree earned/year/contiguous; native & other languages, occupation, ethnicity, smoker, height, weight, city, state, country born & raised for subject, mother & father: all self reported
- https://catalog.ldc.upenn.edu/LDC2013S03

Enhancing Found Data

- Corpus: collection of recordings of linguistic behavior selected, and possibly annotated, for a specific purpose.
  - reuse generally requires re-annotation and possible re-sampling
- Differences = Challenges = Opportunities
  - domain of inquiry (e.g. versus speech community)
  - model of the phenomena (feature, variable) under study
  - sampling (talkers, tokens)
  - metadata

Linguistic Data Consortium

Panel Corpora on a shoestring budget

Old/New Wave of Panel Corpora

As studies of aging queens show

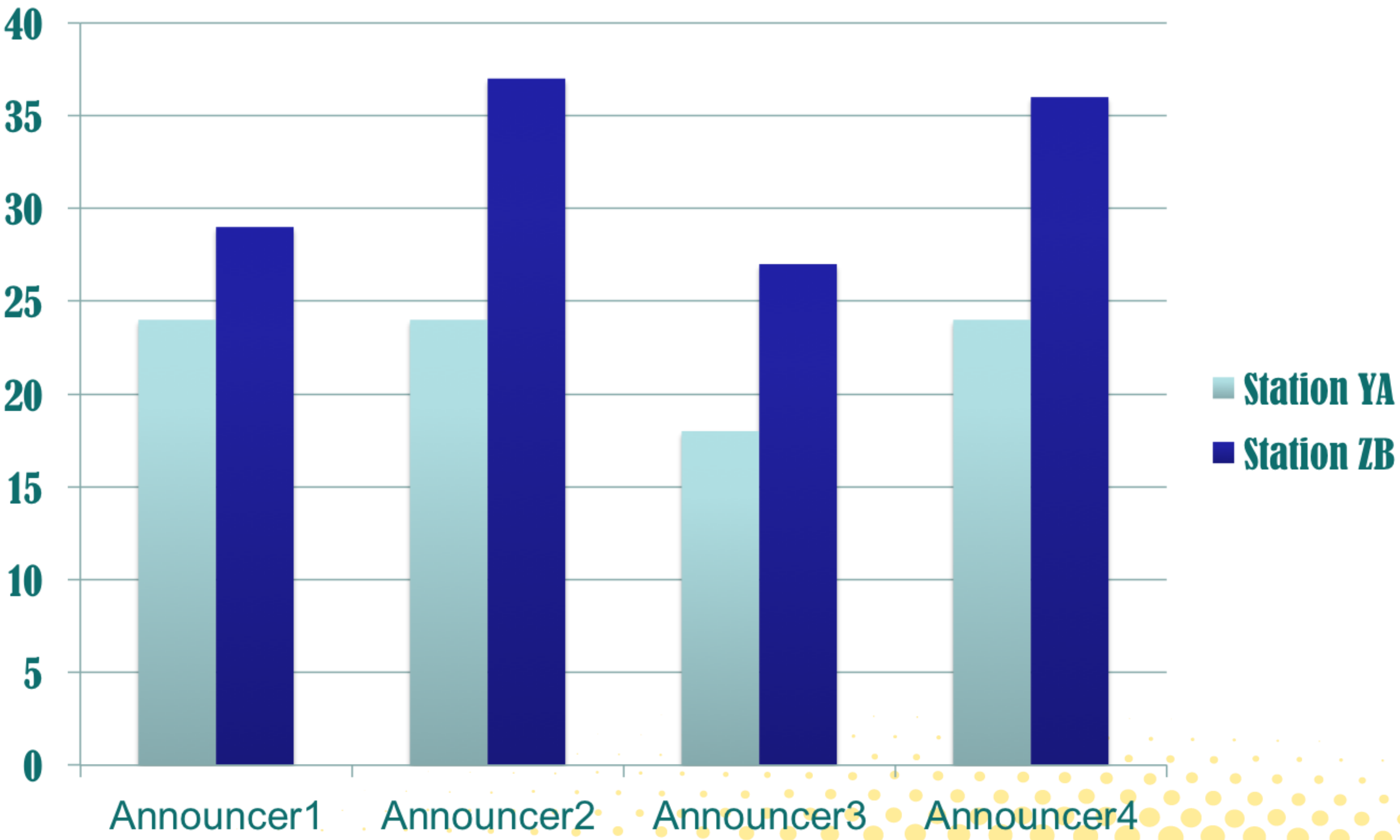
- Studies by Harrington (Elizabeth II) & Quené (Beatrix) show

- Even queens change over time, with no obvious motivation

i. Many Performers

have several avatars

Bell (1984:171)



Singers like Bob Dylan

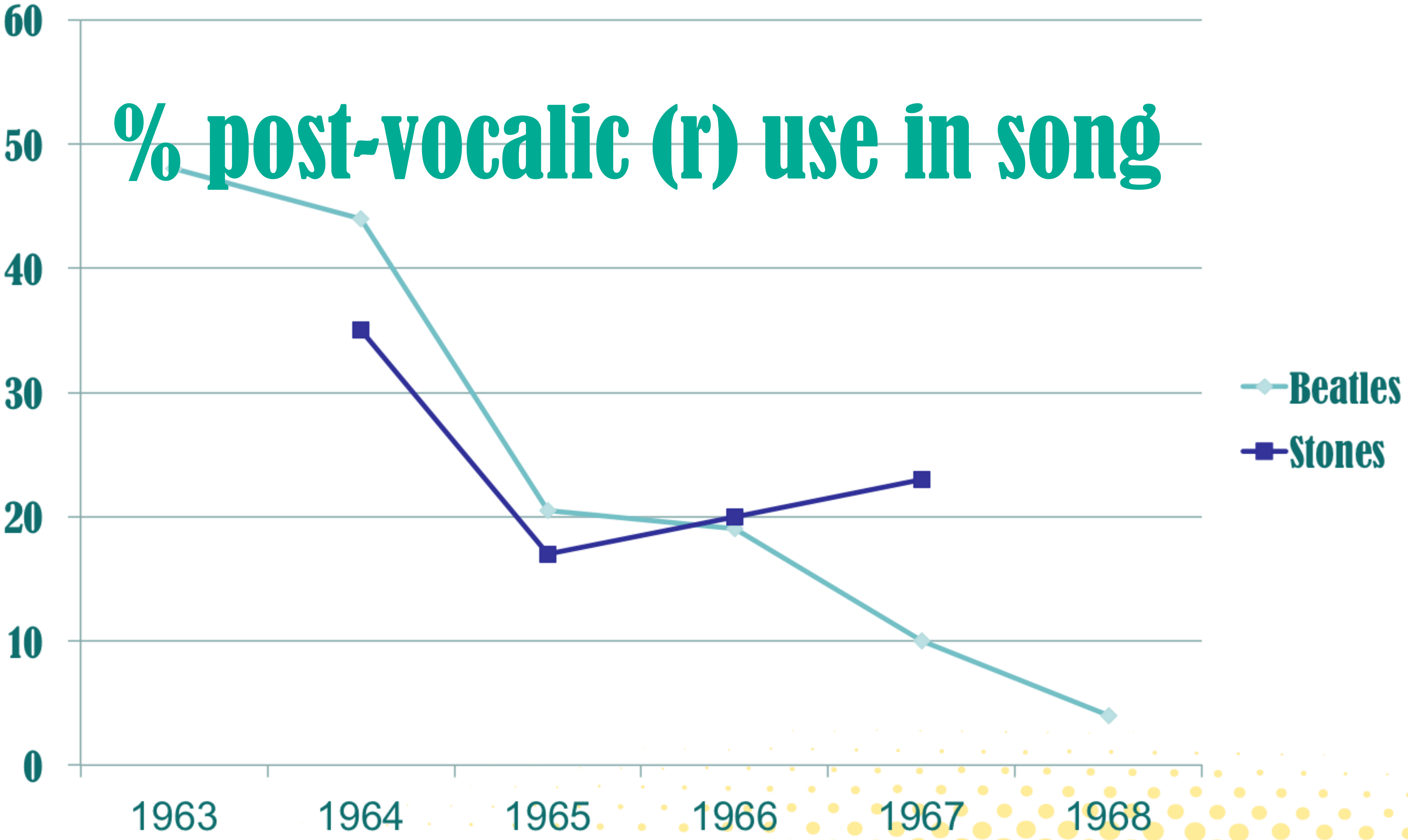
- Guthrie
Christian Rt
Hassidic Jew
Cowboy
Sr. Citizen

Or not

- But even when they don't
- We can find that they change over time (e.g., Trudgill 1983)…

Trudgill (1983:152)

% post-vocalic (r) use in song



Not merely ‘real time’

Many Politicians

SHUCKS, I KNOW A DODDERING OLD MOVIE ACTOR LIKE ME IS NO MATCH FOR A YOUNG CARD SHARK LIKE YOURSELF, AND MY MIND WANDERS AND, GIN!... I'M SLOW, AND...

have several avatars

Politicians like: LBJ--

Situational factors are critical.

e.g., LBJ

The novice

Sam Rayburn’s southern boychik

Southern Senatorial wheeler-dealer

King-Pin of the Senate

‘Demeaned’ Vice President

President

Retiree

LBJ--

Situational factors are critical.

e.g., LBJ

Public vs. private

phone conversations

Who is the interlocutor?

--your best guess on Subject's relationship to the interlocutor?

--your best guess on what is to be accomplished?

Files which are readily available from Miller Center/Library

Backup info on situations available in historical texts.

If the *sound quality* is OK--

This doesn't mean "take anything".

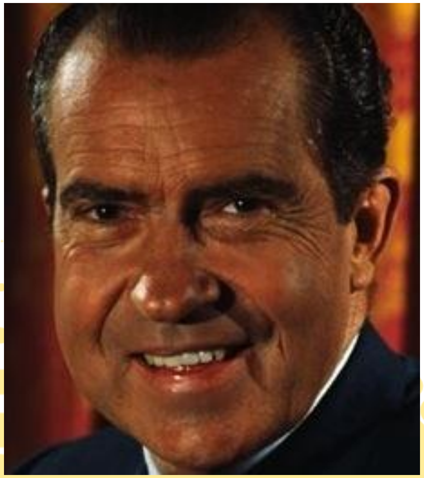
There are aligned sound files

/www.nixontapes.org (by Luke Nichler)

/whitehousetapes.net/transcript/nixon (Miller Center*)

/www.nixonlibrary.gov/forresearchers/find/tapes/watergate/trial/transcripts.php

/www.talkbank.org/data/CABank/Jefferson/



With the BEST transcripts

That nonetheless are useless to us.

A Few More Examples

Bell (1984ff) – NZE Radio Announcers; Advertisers

Coupland (1985f)-Welsh Radio Host

Kemp/Yaeger-Dror (ms) Quebec radio announcers

Clayman/Heritage (1992f) IVs with heads of states

Kemp/Yaeger-Dror (1992-4)-Quebec politicians

Hall-Lew/Yaeger-Dror (2006ff)-American politicos

Yaeger-Dror/Hedberg (2008) US political programs

Hernandez-Campoy (2002f) Murcian political IVs

Soukup (2009f) Austrian political programs

So…what's the take-away msg?

Linguistic Data Consortium

e.g.,

Interactive setting…

- ‘Broadcasts’ are not necessarily the same
  - Bell showed us that even the audience for a radio station influences speech
  - Coupland found the importance of stance
  - How much more so a face to face audience
- We should activate a sense of ‘genre’ and ‘stance’
  - Heritage /Clayman show that even the genre of questioning a head of state varies
  - Yaeger showed ‘news reading’ includes subgenres
  - Yaeger-Dror et al showed that ‘story reading’ varies in interesting ways
  - Campbell-Kibler showed that mood can influence attitude twd interlocutor

-Panel studies should simplify – maintain narrow focus

- Or vary systematically

e.g.,

We now are learning to vary the TIME dimension…

Future panel studies can systematically alter others

As they are doing with the latest LDC Philadelphia community studies.

We should also interpolate an understanding of a

‘nested set of repertoires’ from earlier studies

Public Speakers

- Their sound files don't die, but become more accessible
- Often with transcripts aligned – by Miller Center or other
- Radio Announcers (Bell 1984)
- - Different stations/ social settings
- - Different program genres
- - Change over time
- E.g, if you're seeing how sports broadcasters talk, analyze 1 setting
- Use appropriate backup—even wikipedia helps
- - e.g., Politicians (MYD, Clayman & Heritage, LHL & MYD, Hall-Lew et al…)
- - who they are relative to a given interlocutor on that day
- - social setting

Other Issues

- **Beware changing attitudes /conventions**
- - Trudgill's singers only changed their attitude twd British dialects
- - CBF sports' heroes/broadcasters changed attitudes twd MF
- **Integrate multiple social variables into study**
- - '*Bricolage*' - different variables presented for different identities
- Often 'nested',
- Often with separate variables for different identities (Negron 2014)
- E.g., Becker (2013), Cutler (2010) – AAE, Latino,…& NYC
- E.g., Yaeger-Dror (fc)—*Mizraxi*/MidEastern…& Israeli
- E.g., Hall-Lew et al (2013)—TeaParty/Rancher & Locality
- Interpolate what's learned in 1 study into the archival record!?

Linguistic Data Consortium

Lessons from those better funded

- Get enough people in your sample #1 that you can afford to lose some on the way to the reIV.
- - of the 120 speakers of MF in 1971,
  - only 60 were still available for the 1984 interviews.
- Train IVers, so the second IV setting mimics the first
- -MF IVers were about the same demographic as the first
- -and listened to those IVs to adapt their presentation
- Beware changing attitudes /conventions
- Integrate multiple variables into study
- [cf. Thibault & Daveluy, Gregersen's plenary on Monday]

Lessons from those better funded

- {BNC} cannot/does not provide possible panel data
- - because we can’t get sufficient background information,
- How can we incorporate information that we want salvaged?

Linguistic Data Consortium

Thanks to …

- The organizers
- Linda Brandschain & Chris Cieri,
- Cece Cutler & Andy Gibson
- Howie Giles, Dave Graff, Greg Guy,
- Lauren Hall-Lew, Jake Harwood
- Tyler Kendall,
- LDC
- Carmen Llamas & Dom Watt
- Richard Ogden, Brendan O'Connor
- Jane Stuart-Smith
- for many invigorating discussions!
- You for your interest!
- & Good Luck with *your* work