

# The Sociolinguistic Archive and Analysis Project: Data, Tools, and Applications

Tyler Kendall  
University of Oregon  
tsk@uoregon.edu



LDC Institute, May 18 2012

# Agenda

- An introduction to the Sociolinguistic Archive and Analysis Project (SLAAP)
- A **demonstration** of some of the project's software features
- A little about how the software works
- An application from some current research
  - Using speech timing to model sociolinguistic variables



# Background: The NCLLP

- The North Carolina Language and Life Project (NCLLP) is a sociolinguistic research initiative based at North Carolina State University
  - Under the direction of Walt Wolfram
  - Conducts sociolinguistic interviews & studies with a wide array of individuals and groups throughout North Carolina and elsewhere (predominately in the American South)
  - Has one of the largest collections of sociolinguistic interview recordings in the U.S.
    - > 2,500 interviews (and growing)
  - In early 2005, with support from the NCSU Libraries, we began a formal digitization and organization initiative for our large media collection

# The Sociolinguistic Archive and Analysis Project (SLAAP)

- On the one hand, we have been organizing (and digitizing) the sociolinguistic archive collection of the NCLLP, and increasingly others, for preservation and accessibility
  - Making the collection web accessible, so scholars can access their data from anywhere in the world
- But SLAAP is more than an archive:
  - It is web-based software that seeks to enhance sociolinguistic data through the development of analytic tools and data-models
  - Through this, we are exploring new, computer-enhanced techniques for interacting with the collection and for conducting sociolinguistic analyses
  - In Poplack's (2007) terms, SLAAP is a **tool** with no projected **end-product**
  - It is an example of a more explicit conception of data in sociolinguistic practice (Kendall 2008)

# The (Current) Archive

- Currently (May 2012), SLAAP houses:
  - Over 2,600 interviews
  - Over 4,100 media files (> 2,100 hours of audio)
  - Not just from North Carolina...
    - Over 68 hours of time-aligned transcripts (~700,000 words)
      - ~27 hours and growing from the West Virginia Dialect Project (PI Kirk Hazen, WVU)
      - ~7 hours from South Texas (PI Erik Thomas, NCLLP)
      - ~12 hours from Washington DC (Mallinson & Kendall 2009)
- But SLAAP seeks to be more than just an archive...

# Demo: <http://ncslaap.lib.ncsu.edu/>

**1** main library view

**2** full record view

**3** listen and annotate

**4** download and extract

**5** variable tabulation tool

**6** transcript features

**7** speaker-pitch analysis

NC SLAAP Archive: Browse | Search

[View: Long | Short] | Site: princeville | Showing 5 (of 5) records

Interview	Site	Speaker(s)	Interview Info	Media	Transcript
prv007a	Princeville PEO	black female, born 1964	Date: 09/26/2003 Interviewer(s): RR, DG Contains: sociolinguistic interview	prv007aa [Listen] [Download] prv007ab [Listen] [Download]	prv007aa_1985_2090 prv007aa_...
prv007b	Princeville PEO	black female, born 1964	Date: 09/26/2003 Interviewer(s): RR, DG Contains: car tour of town	prv007ba [Listen] [Download] prv007bb [Listen] [Download]	
prv0110a	Princeville SK	black male, age 55	Date: 10/03/2003 Interviewer(s): RR, DG Contains: sociolinguistic interview, ?	prv0110a [Listen] [Download] prv0110b [Listen] [Download]	prv0110a_...
prv0110c	Princeville SK	black male, age 55	Date: 02/21/2005 Interviewer(s): RJ Contains: radio interview	pvlv0111f [Listen] [Download] pvlv0112f [Listen] [Download] pvlv0113f [Listen] [Download] pvlv0114f [Listen] [Download] pvlv0115f [Listen] [Download] pvlv0116f [Listen] [Download] pvlv0117f [Listen] [Download]	pvlv0115f_5 pvlv0115f_8
prv021v	Princeville PEO	black female, born 1964	Date: 02/18/2005 Interviewer(s): DG	pvlv021v [Listen] [Download] pvlv021v [Listen] [Download] pvlv021v [Listen] [Download]	pvlv021v_5

LP Staff Tools | Library Access

Please delete some files or enable automatic cleanup.

[Tabulation Summary] | [Transcript Summary] | [Speaker Analysis] | [Manage Sound]

[Search Annotations]

Audio Interface

Audio File: [Choose File] [Browse] [Cancel] [Upload]

Annotations: [Add Annotation] [Remove Annotation] [Clear Annotations]

Length: 45:42 min (-2:185.7 sec)

Download complete file | Extract segment from audio file

Start time (sec): [0:00] [OK] [Cancel]

End time (sec): [0:00] [OK] [Cancel]

Start time (sec): [0:00] [OK] [Cancel]

End time (sec): [0:00] [OK] [Cancel]

Speaker-pitch analysis

Speaker: [Choose Speaker] [OK] [Cancel]

Start time (sec): [0:00] [OK] [Cancel]

End time (sec): [0:00] [OK] [Cancel]

Exporting 100 lines (with columns between 1 and 2 seconds) starting 1 line.

transcript features

00:00:00 [0:00] in back seat

00:00:01 [0:01] and (part) of the people in back

00:00:02 [0:02] (part)

00:00:03 [0:03] (part)

00:00:04 [0:04] (part)

00:00:05 [0:05] (part)

00:00:06 [0:06] (part)

00:00:07 [0:07] (part)

00:00:08 [0:08] (part)

00:00:09 [0:09] (part)

00:00:10 [0:10] (part)

00:00:11 [0:11] (part)

00:00:12 [0:12] (part)

00:00:13 [0:13] (part)

00:00:14 [0:14] (part)

00:00:15 [0:15] (part)

00:00:16 [0:16] (part)

00:00:17 [0:17] (part)

00:00:18 [0:18] (part)

00:00:19 [0:19] (part)

00:00:20 [0:20] (part)

00:00:21 [0:21] (part)

00:00:22 [0:22] (part)

00:00:23 [0:23] (part)

00:00:24 [0:24] (part)

00:00:25 [0:25] (part)

00:00:26 [0:26] (part)

00:00:27 [0:27] (part)

00:00:28 [0:28] (part)

00:00:29 [0:29] (part)

00:00:30 [0:30] (part)

00:00:31 [0:31] (part)

00:00:32 [0:32] (part)

00:00:33 [0:33] (part)

00:00:34 [0:34] (part)

00:00:35 [0:35] (part)

00:00:36 [0:36] (part)

00:00:37 [0:37] (part)

00:00:38 [0:38] (part)

00:00:39 [0:39] (part)

00:00:40 [0:40] (part)

00:00:41 [0:41] (part)

00:00:42 [0:42] (part)

00:00:43 [0:43] (part)

00:00:44 [0:44] (part)

00:00:45 [0:45] (part)

00:00:46 [0:46] (part)

00:00:47 [0:47] (part)

00:00:48 [0:48] (part)

00:00:49 [0:49] (part)

00:00:50 [0:50] (part)

00:00:51 [0:51] (part)

00:00:52 [0:52] (part)

00:00:53 [0:53] (part)

00:00:54 [0:54] (part)

00:00:55 [0:55] (part)

00:00:56 [0:56] (part)

00:00:57 [0:57] (part)

00:00:58 [0:58] (part)

00:00:59 [0:59] (part)

00:01:00 [1:00] (part)

Assorted screenshots from SLAAP, from Kendall (2007)

# How SLAAP Works...

# SLAAP's data model

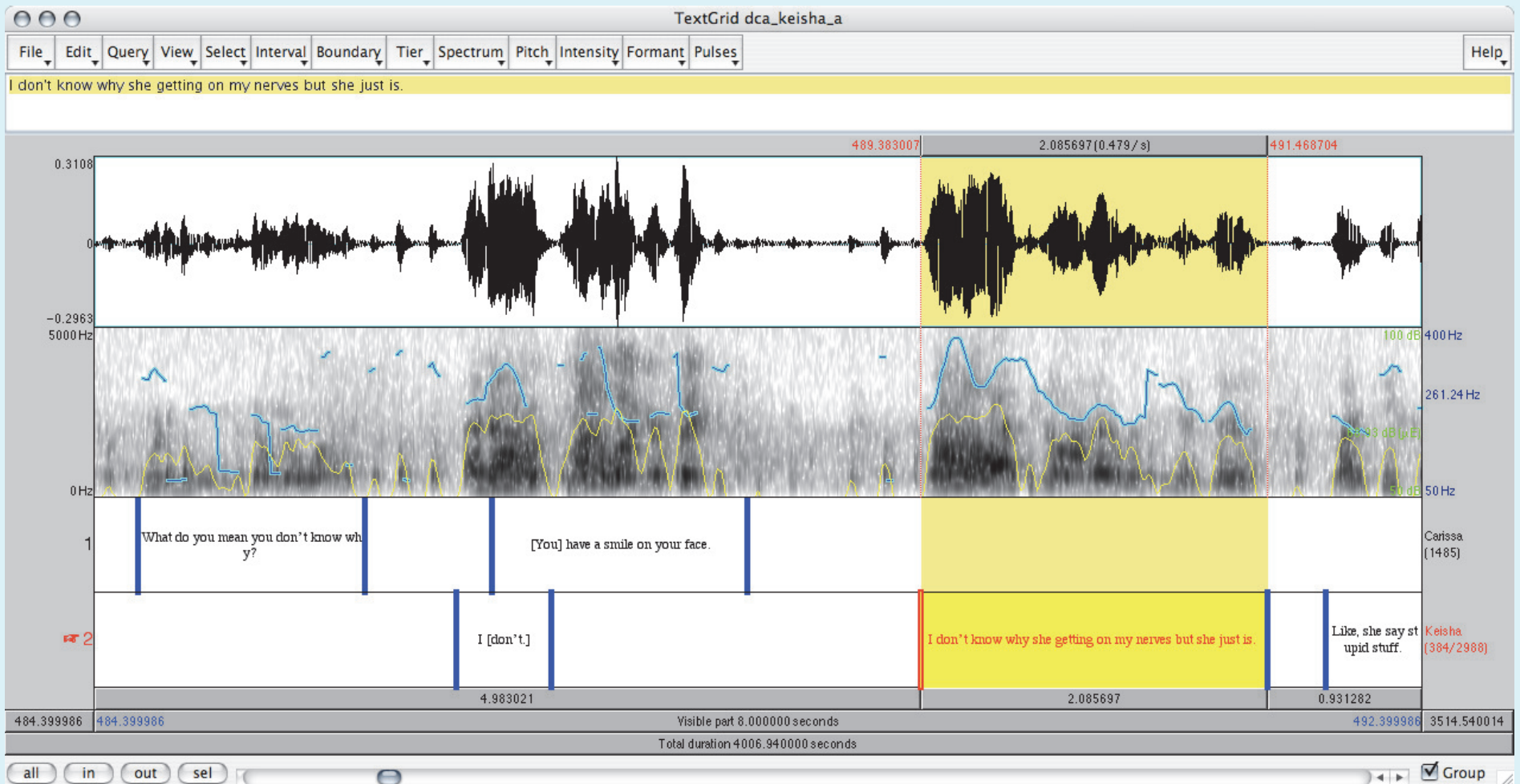
- Media files are most basic level of “data”
  - They are housed in file space, but their locations and information about them are stored in a database
- Projects, interviews, speakers, etc. are housed in a relational (MySQL) database
- Information like notes (“annotations”), variable tabulations, transcripts, etc. are time-stamped entries also in database tables

E.g., transcripts are comprised of data-base entries:

Start Time	Speaker Reference	Simple Orthographic Representation	End Time	Additional (meta)data
------------	-------------------	------------------------------------	----------	-----------------------



# Transcripts Generated in Praat

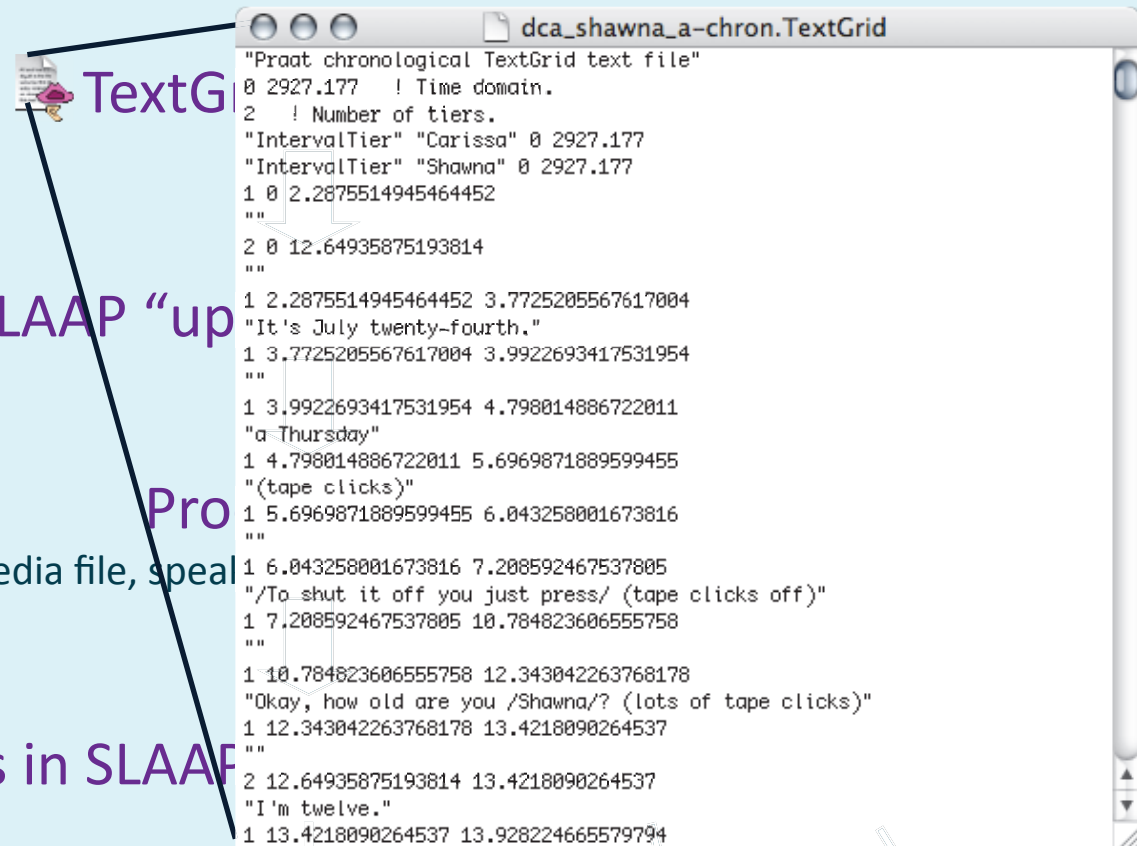


(media: dca\_keisha\_a)

# What constitutes a transcript line?

- The phonetic utterance
  - The “hypothesis” behind SLAAP is that a powerful **methodological** definition of the utterance is:
    - Silence - **sound** - silence on the part of a speaker
  - All silence > ~60 ms is marked as separate from speech
- nb.
  - In this approach: Timing information is central. The text is an *approximation* of the speech, more a link to the audio than a representation of it.
  - ~ Annotation Graphs (Bird & Liberman 2001)

# Adding a Transcript to the Archive



```
"Praat chronological TextGrid text file"
0 2927.177 ! Time domain.
2 ! Number of tiers.
"IntervalTier" "Carissa" 0 2927.177
"IntervalTier" "Shawna" 0 2927.177
1 0 2.2875514945464452
""
2 0 12.64935875193814
""
1 2.2875514945464452 3.7725205567617004
"It's July twenty-fourth."
1 3.7725205567617004 3.9922693417531954
""
1 3.9922693417531954 4.798014886722011
"a Thursday"
1 4.798014886722011 5.6969871889599455
"(tape clicks)"
1 5.6969871889599455 6.043258001673816
""
1 6.043258001673816 7.208592467537805
"/To shut it off you just press/ (tape clicks off)"
1 7.208592467537805 10.784823606555758
""
1 10.784823606555758 12.343042263768178
"Okay, how old are you /Shawna/? (lots of tape clicks)"
1 12.343042263768178 13.4218090264537
""
2 12.64935875193814 13.4218090264537
"I'm twelve."
1 13.4218090264537 13.928224665579794
```

SLAAP “up

Pro

(e.g., linked to media file, speak)

Resides in SLAAF

SLAAP

can process  
in a variety of ways

Can script  
with Praat  
or R or ...

Can manipulate  
directly in MySQL

Can be exported  
in “any” format  
(e.g., XML)

Etc.

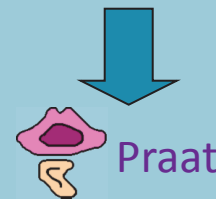
\* Also, for non-SLAAP users, a public tool to convert TextGrids to plain text on the website

# SLAAP itself works by...


- All web pages are generated by PHP scripts
- Acoustic analysis features work by:
  - Extracting relevant information from the database (timestamps, media file location, etc.)
  - Sending this information to Praat via customized Praat scripts
  - Reading Praat output (and optionally post-processing, via software like ImageMagick & LAME)
  - Compiling and formatting this output and sending it back to the user.

- E.g., extracting pitch info:

Start Time	Spkr Ref	Simple Ortho Rep	End Time
858.926	PEO	I became a commissioner	860.049



```
./Praat_4_3_12_exe praat_scripts/make_excerpt_pitchtier.praat prv007aa  
soundfiles/ praat_out/tskendal/ 858.925932 860.049443 75 600 0.01
```

Line Start	Spkr	Pitch & Text	End
			
<23>	[858.926] PEO:	I became a commissioner	[860.049]

# An Application...

Using speech timing to investigate  
language variation

# Sequential temporal patterns of speech & The Henderson graph

- On account of the finely time-aligned transcripts, SLAAP is well-suited for investigating speech timing phenomena (Kendall 2009, forthcoming)
- Here, focusing on just one idea within variationist sociolinguistics (Labov 1966, 1972, ...)
  - Channel cues to attention to speech, via a visualization technique that, following Levelt (1989), I call a **Henderson graph** (Henderson, Goldman-Eisler, and Skarbek 1966)

# Attention to speech

- William Labov's foundational work in variationist sociolinguistics (1966, 1972) established that a speaker's **attention to** his/her **speech** is an important factor in his/her language realization.

Within-speaker variation

→ **Style**

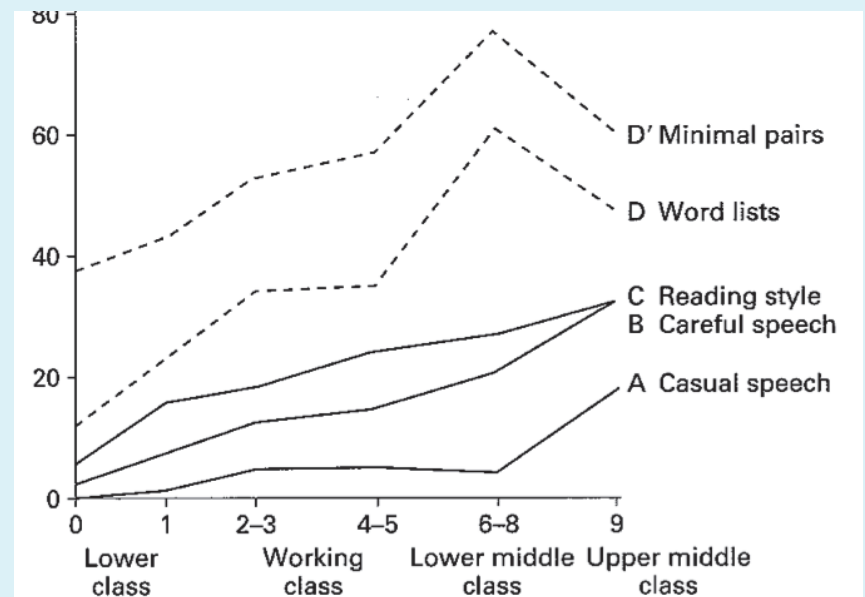


Figure 7.10 Simplified style stratification of (r): six class groups

Labov 1966/2006: pp. 150-151

# Channel cues

- In his formulation of the attention to speech model, Labov (1966, 1972) proposed that **channel cues** can indicate changes in attention to speech
  - To be sure that the different interview tasks succeed in eliciting different styles, it was necessary to look to other aspects of the sociolinguistic interview data (i.e., other aspects of the speech recording)
- **Channel cues = pauses, laughter, breathing, ...**



## Labov 1972: 94-95

- *It is of course not enough to set a particular context in order to observe casual speech. ... **The best cues are channel cues:** modulations of the voice production which affect speech as a whole. Our use of this evidence must follow the general procedure of linguistic analysis: the absolute values of tempo, pitch, volume, and breathing may be irrelevant, but contrasting values of these characteristics are cues to a differentiation of Style A and Style B. **A change in tempo, a change in pitch range, a change in volume or rate of breathing, form socially significant signs of shift towards a more spontaneous or more casual style of speech.***

# But,

- The idea was tabled because no systematic way to actualize the idea
  - *“It appears that channel cues did not provide a high enough level of interpersonal reliability for most researchers” (Labov 2006: 74).*
- E.g., Wolfram (1969: 58-59)
  - *“An exploratory attempt to distinguish careful from casual speech based on Labov’s criteria was rejected for several reasons. In the first place, any of the paralinguistic channel cues cited as indications of casual speech can also be indications that the informant feels an increased awareness of the artificiality or formality of the interview situation. Can nervous laughter reliably be distinguished from casual or relaxed laughter? Also, **the subjective interpretation of the paralinguistic cues tends to bias the interpretation of casual speech ... To what extent must there be a change of pitch or rhythm and how close to the actual feature being tabulated must it occur?**”*

# New old questions

- I have been interested in drawing on the fact that **attention** is a more broadly studied phenomena
- And that others have thought about the role of channel cues – like **pauses** – in speech production
- Can we revive the channel cue component?
  - Can we quantify it and use attention to speech in new ways?
  - In Kendall (2009, forthcoming) I draw on work by Frieda Goldman-Eisler and other psycholinguists who have studied pause and hesitation phenomena
    - And also, e.g., Wallace Chafe’s (1994) ideas about the “flow of consciousness” in discourse

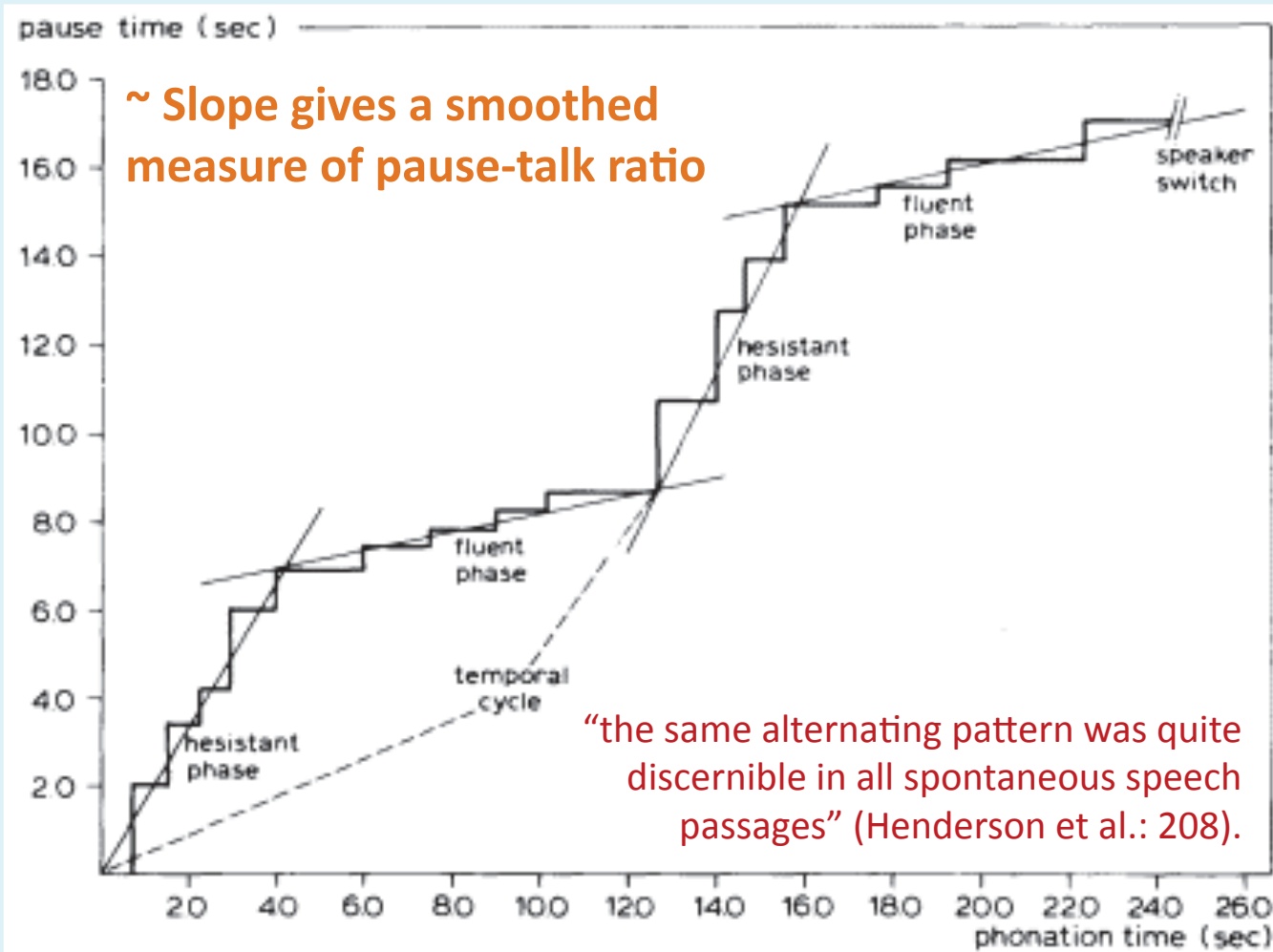
# Goldman-Eisler and pauses

- Goldman-Eisler (e.g., 1968), and others, showed that pauses are a cue to language production processes
  - E.g., that pauses are more likely and longer before words with less predictability and with more difficult tasks... that pauses can be used “to sort out which parts of verbal sequences are verbal habits and which are being created at the time of speaking” (1968:43).
- There is “a lawful relationship between temporal phenomena in human speech and concurrent cognitive processes” (Kowal and O’Connell, 1980:61).

# The Henderson graph

(Henderson, Goldman-Eisler, and Skarbek 1966)

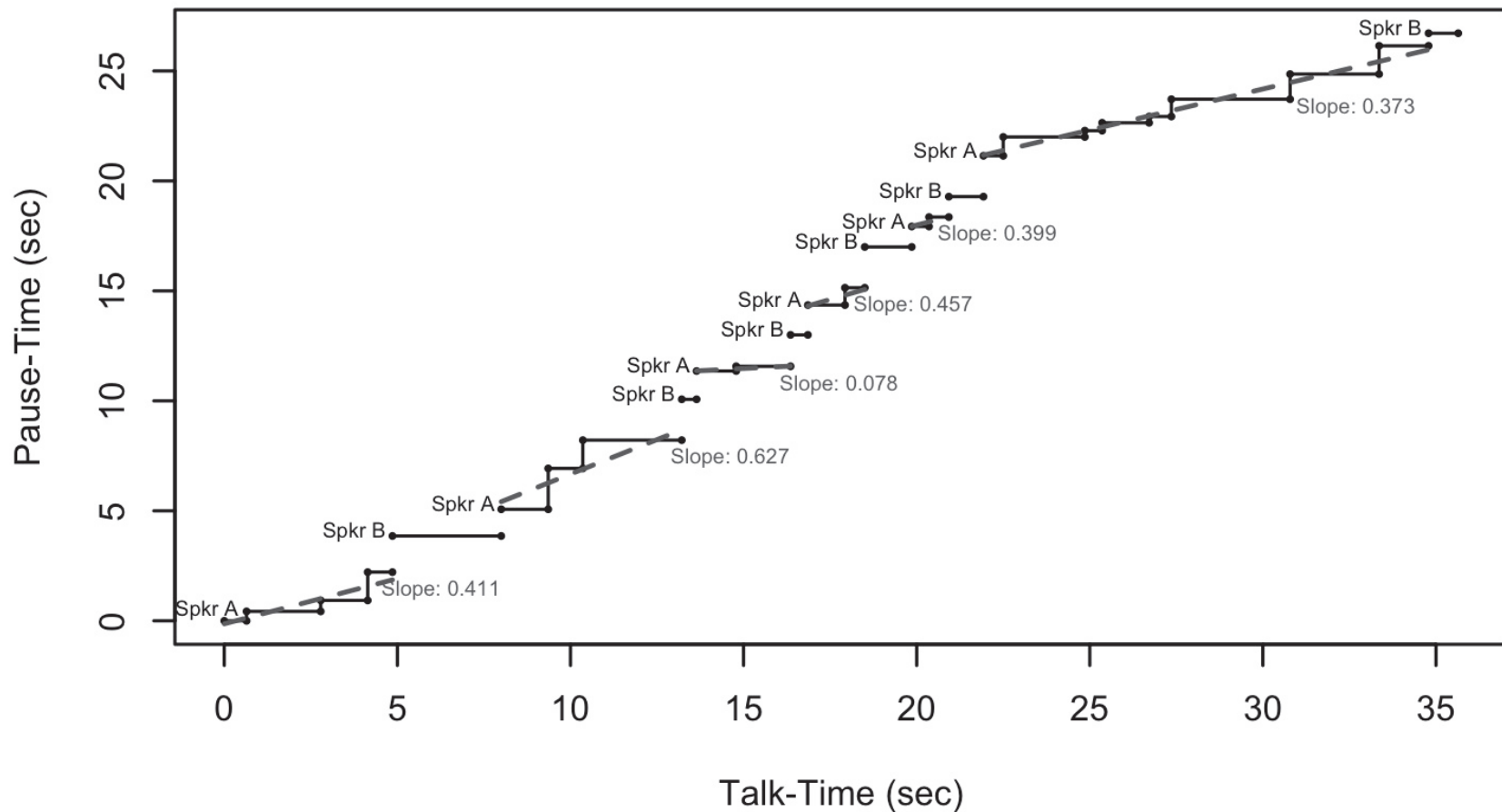
Levelt 1989: 127, Fig. 4.3



The Henderson graph is a representation of a speech event in which talk-time is plotted on the x-axis while pause-time extends on the y-axis. Changes in the characterization of the talk are viewable as changes in the slopes of sections of the talk.

# A Henderson graph of interview talk

Henderson Graph example dialogue



# Henderson graphs in SLAAP

Transcript - w\_dca\_alayna\_a\_0\_4245

http://ncslaap.lib.ncsu.edu/ncslaap/transcript.php?t=w\_dca\_alayna\_a\_0\_4245&format=paragraphs&pre\_resize=400

NC STATE UNIVERSITY [ User Forum ] [ tskendal : Acct | Logout ]  
(autologout if no activity at 11:52:53)

[ Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools ] [ @transcript.php ]  
[ Library ]

SLAAP v. 0.95 - Transcript - w\_dca\_alayna\_a\_0\_4245

Disk Usage: 370,444 kb Large Disk Usage: Please delete some files or enable automatic cleanup.

Transcript: dca\_alayna\_a\_0\_4245 [ Interview: dca\_alayn ] (Sorry: Not all the options on this page are currently working.) [ All Transcripts ]

[ View Extra Information: No Summary | Notes  ] [ Show Audio Player  ] [ Auto-Summarize | Export ]

[ Options: Hide Line#s  | Hide Pauses/Blank Lines  | Show Gaps  | Hide Times  | Indent Overlap  ] [ Display: Paragraphs ]

[ Transcript Window Size: 600 px ] [ Show All | 4665 Lines | Start at Line 1 ] [ Show Annotations Inline  | Edit Links?  | No Tab Links ]

Audio: Play (p) Stop (s) [ Get Time | Move to Time (m) : CURSOR at 274.28 SEC | << Annotate at Time ]

Carissa: <sup>372(au)</sup>Okay. [pause 0.21] <sup>374(au)</sup>And then before that, how long were you in [pause 0.07] <sup>376(au)</sup>foster [pause 0.52] <sup>378(au)</sup>care home? [gap 0.03]

Alayna: <sup>380(au)</sup>Three months, or [pause 0.23] <sup>382(au)</sup>four months. [gap 0.25]

Carissa: <sup>384(au)</sup>Did you move around to different houses, were you in one [house?]

Alayna: <sup>385(au)</sup>[I] stayed in one h-, foster home [gap 0.06]

Carissa: <sup>388(au)</sup>Okay. [pause 1.61] <sup>390(au)</sup>Okay. [pause 0.43] <sup>392(au)</sup>So what's your relationship like with [pause 0.21] <sup>394(au)</sup>Danielle's mom? [gap 0.50]

Alayna: <sup>398(au)</sup>I mean, it's like, a mother and daughter relationship like I had with [pause 0.46] <sup>398(au)</sup>uh, my mom. [gap 0.27]

Carissa: <sup>400(au)</sup>Uh huh. [gap 0.88]

Alayna: <sup>402(au)</sup>So my mom, she's a very, [pause 0.38] <sup>404(au)</sup>she's [pause 0.17] <sup>406(au)</sup>getting what she needs now, my mom. And she carries herself better. [pause 1.14] <sup>408(au)</sup>Cause she uh [pause 0.35] <sup>410(au)</sup>She's doing better than what she is, I'm proud of my mother [pause 0.42] <sup>412(au)</sup>for that. [pause 0.80] <sup>414(au)</sup>So. [gap 0.08]

Carissa: <sup>416(au)</sup>Has she gone through some treatment [pause 0.23] <sup>418(au)</sup>stuff? [gap 0.51]

Alayna: <sup>420(au)</sup>[/I mean/ she, she's,]

Carissa: <sup>421(au)</sup>[/unintelligible/] [gap 0.25]

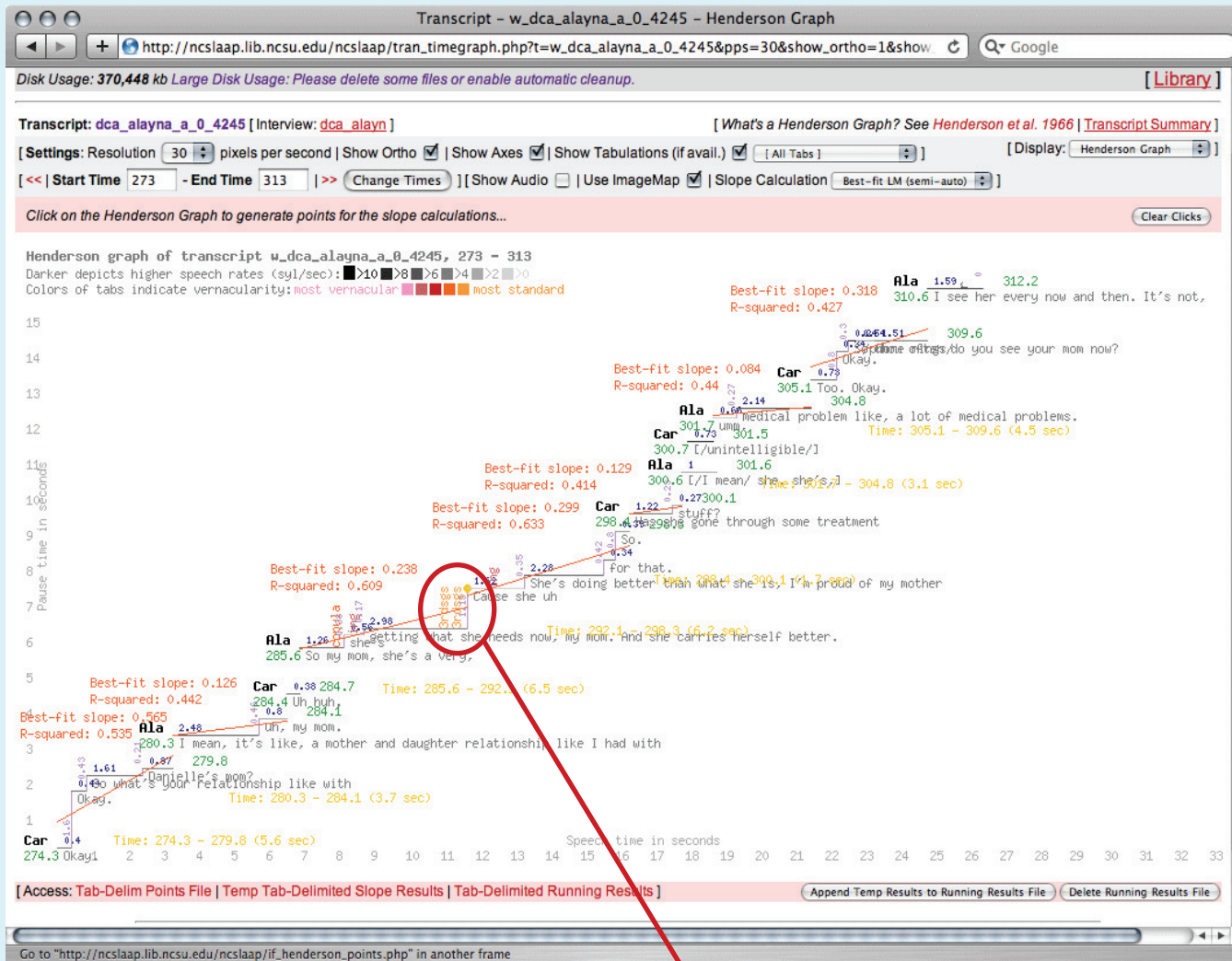
Alayna: <sup>424(au)</sup>umm, [pause 0.27] <sup>426(au)</sup>medical problem like, a lot of medical problems. [gap 0.35]

Canceled opening the page









Coded variables appear in situ

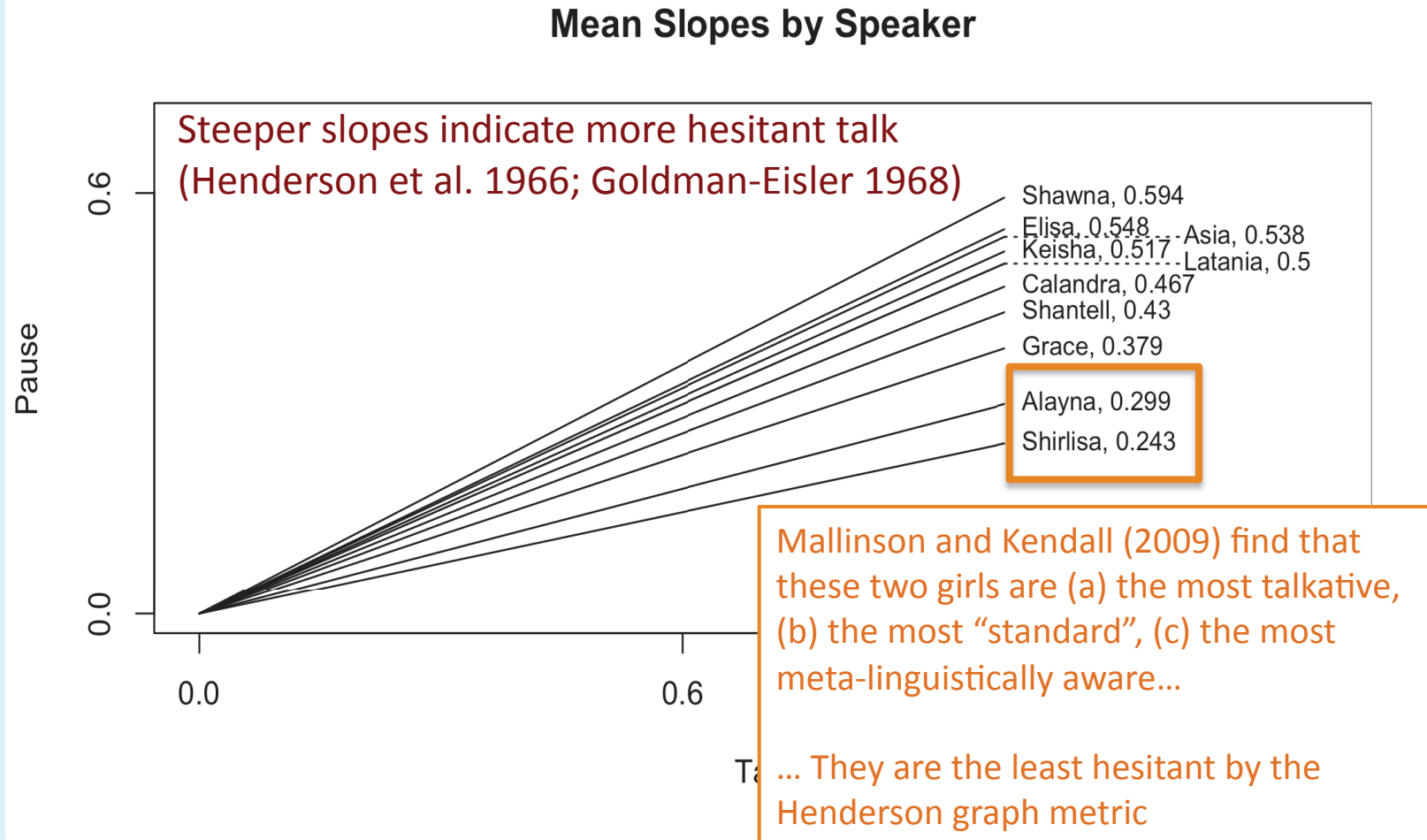
# Sequential temporal variables

- We can envision (and test) a range of predictors
- Related directly to the slope measurements
  - **Slope** = Best-fit slope
  - **$\Delta$ Slope** = Change in slope from previous Henderson segment
  - ... (E.g., **SlopeComp** = Comparison of a given slope and that speaker's mean slope (low, norm, high))
- Also, related to the segmentation made by the slopes
  - **ArtRate** = Median articulation rate within segment ( $\sigma$  / second, not including pauses)
  - **SpkRate** = Overall speaking rate for segment (total #  $\sigma$  / duration of total segment)
  - ... (E.g., Duration, Number of Words, etc.)

# Case Study: African American teens in Washington, DC

- Mallinson and Kendall (2009) examined ten interviews with inner-city, “at risk” African American adolescent girls conducted by their camp counselor, a white female sociologist from Minnesota
  - Semi-structured counseling interviews centered on teens’ home and social lives, gender/sexual ideologies, and aspirations
  - In SLAAP:
    - 10 and 2/3 hours of audio
    - 105,917 words (as transcribed)
- Do we learn anything about these young women and these data through Henderson graph based analyses?

# The slopes themselves



# Sociolinguistic variables

- Mallinson and Kendall (2009; Kendall 2010) examined sociolinguistic variable data for these young women, including:
  - **Copula/auxiliary absence (cop/aux)** with *is* and *are*, as in “She  $\emptyset$  funny” and “They  $\emptyset$  gonna go home”
  - **Velar nasal fronting (ing)**, as in “runnin’” for “running”

Feature	Absent/Total	%Non-Std.
Copula/auxiliary absence, <i>is</i>	185/467	40%
Copula/auxiliary absence, <i>are</i>	267/421	63%
Cop/aux abs, combined	452/888	51%
Velar nasal fronting	1352/1621	83%

# (ing): formality & attention

- Velar nasal fronting, (ing), has been widely found to show stable stylistic and social differentiation in communities
- More attentive speech contains fewer *-in'* forms
  - Steeper HG slopes reflect more hesitant (=more **attentive**) talk...

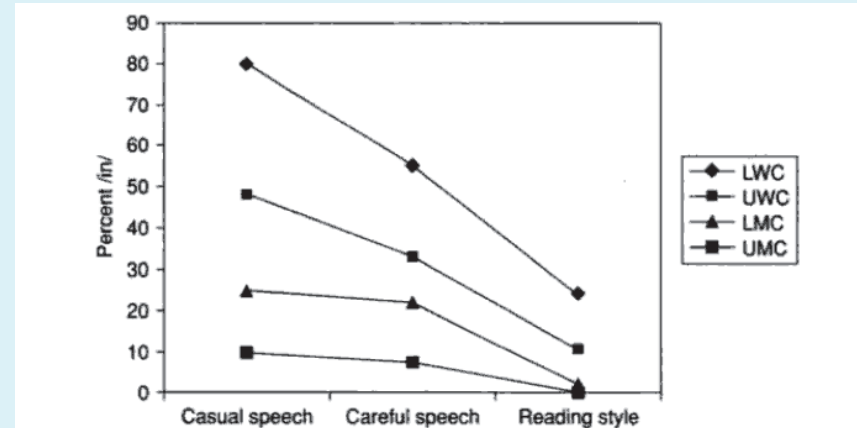


Figure 3.1 Social and stylistic stratification of (ing) in New York City. LWC: lower working class; UWC: upper working class; LMC: lower middle class; UMC: upper middle class (from Labov 1966a).

(Labov 2001)

TABLE 4

FREQUENCY OF -ING AND -IN IN A 10-YEAR-OLD BOY'S SPEECH IN THREE SITUATIONS IN ORDER OF INCREASING INFORMALITY

	-ing	-in
<b>TAT</b>	38	1
Formal interview	33	35
Informal interview	24	41

Chi square: 37.07; P > .001.

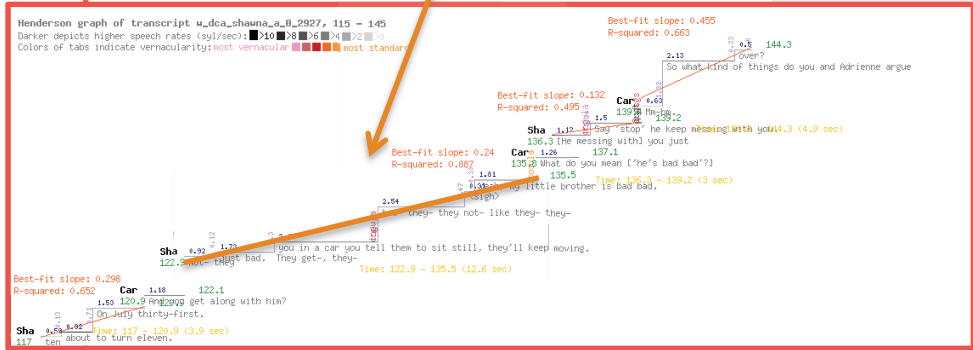
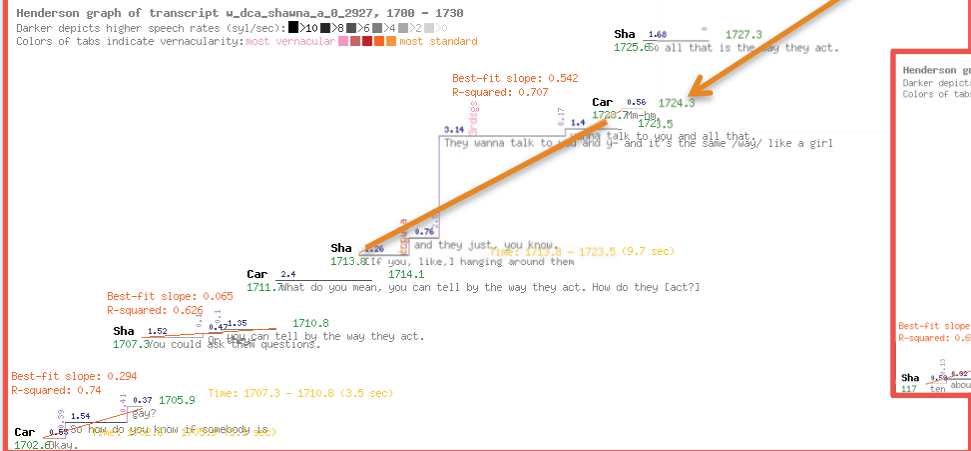
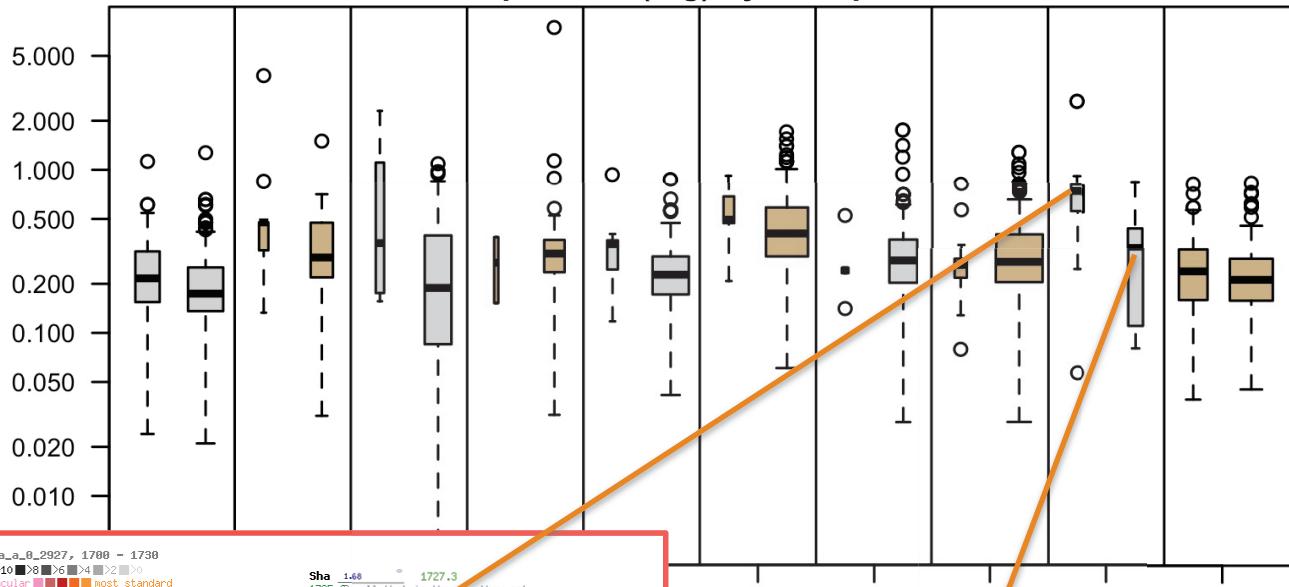
(Fischer 1958)

# Predictions

- Equating attention → hesitancy → steep slopes
- For (ing):
  - From many previous studies, we might expect more full *-ing* realization (and less *-in'*) during steeper slopes than shallower slopes
- For (cop/aux):
  - ? Less clear predictions
    - There is some evidence of stylistic effects on (cop/aux) but less from an attention/formality perspective and conflicting evidence (cf. Rickford and McNair-Knox 1999)
    - No real reason from the literature to expect a correlation with slope (to the same extent as (ing))

# Slopes & (ing)

Slopes and (ing) by DC Speaker

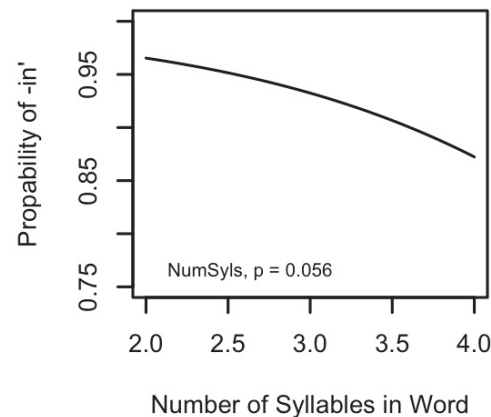
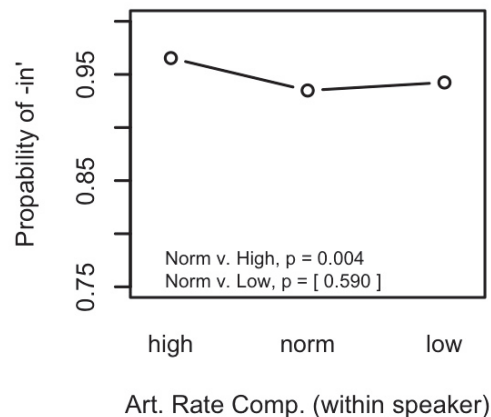
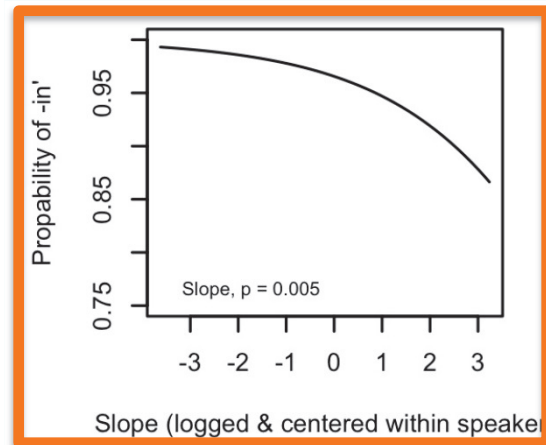
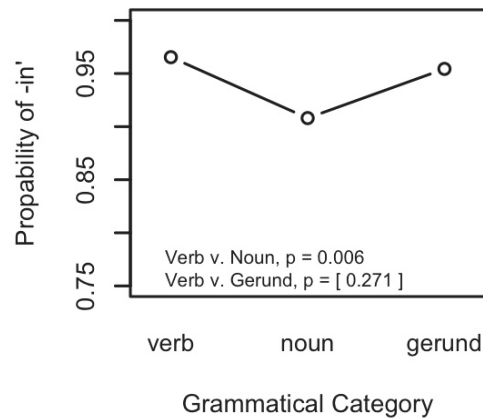


E.g.,  
 “If you, like, **hanging** around them”  
 (slope = 0.54)

E.g.,  
 “...you tell them to sit still they'll keep **movin'**.”  
 (slope = 0.24)

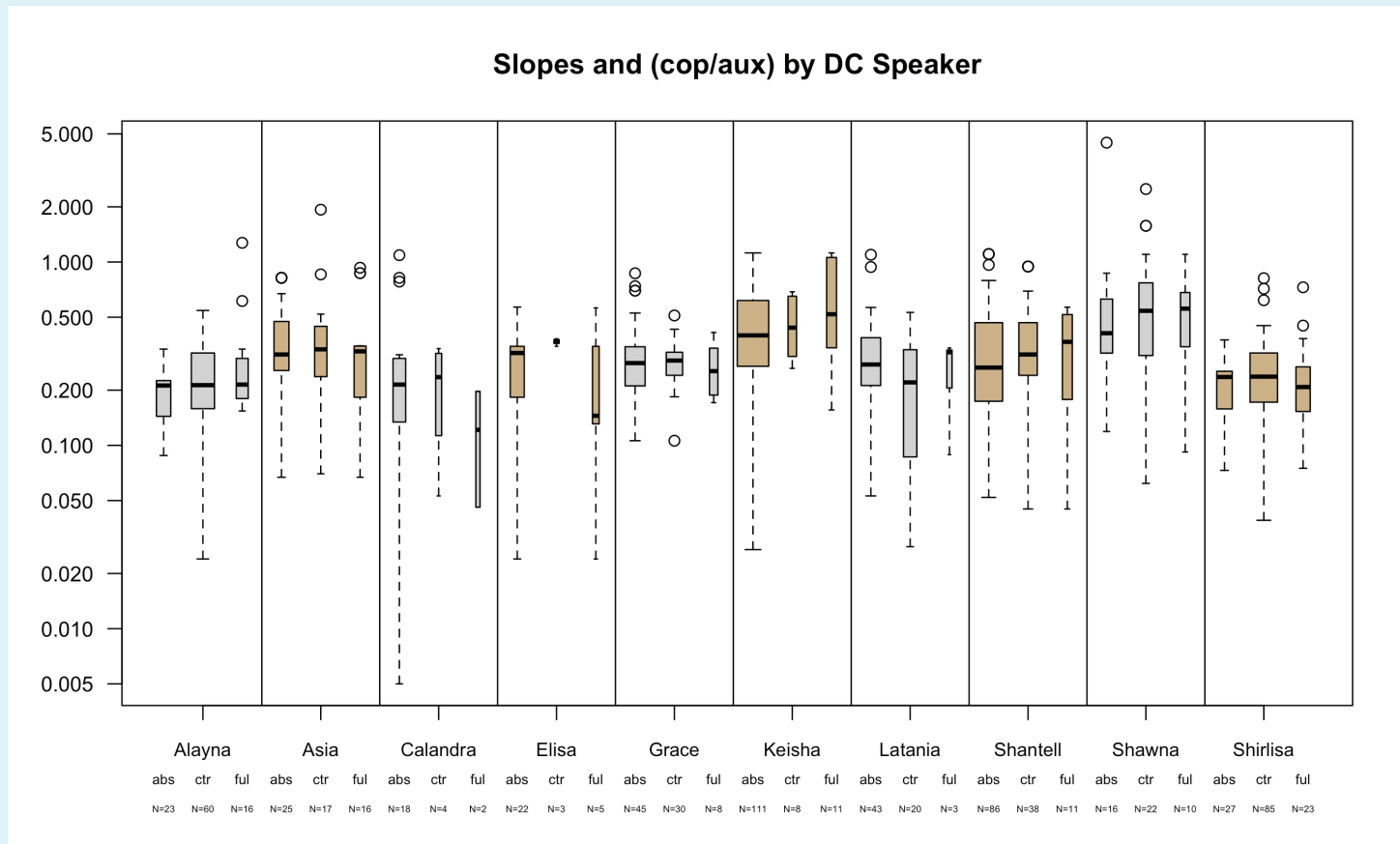


# (ing) statistical results



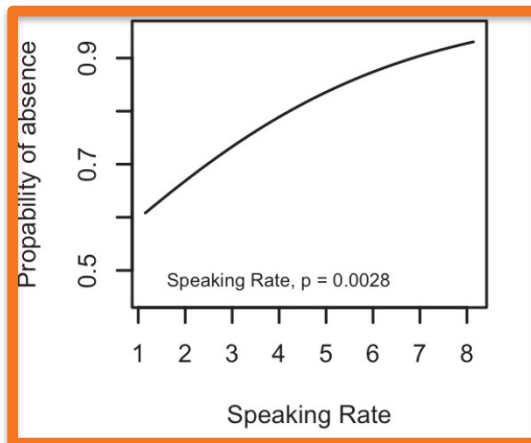
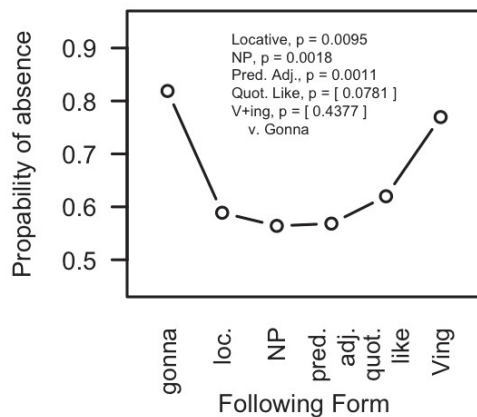
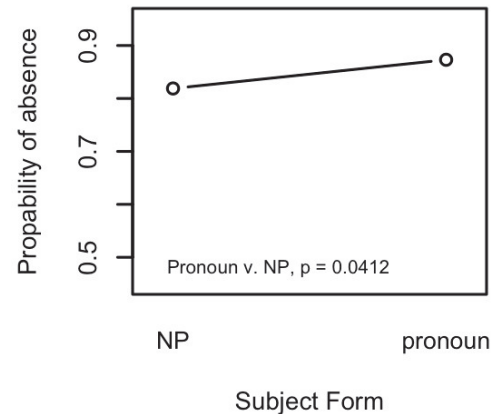
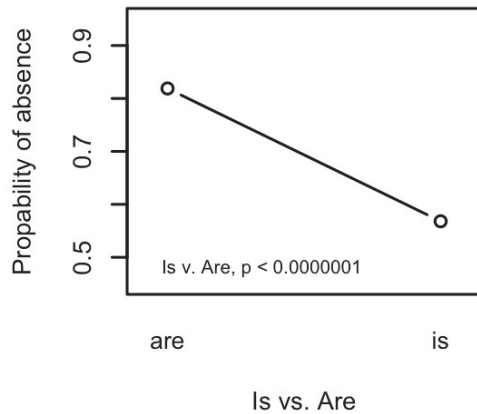
- **Slope  $p = 0.005$** 
  - **Increased slope (more hesitancy) predicts more full -ing!**
  - (Slope is logged and centered on a per speaker basis)
- Also, significant effects of grammatical category, articulation rate, and ~number of syllables in word
- From mixed-effect logistic regression with random intercepts for speaker and word
  - $N = 1,506$

# Copula/auxiliary absence (cop/aux)



- No regular pattern for copula/auxiliary absence
  - Slope is not significant here
- We don't see this effect for every variable...

# (cop/aux) statistical results



- However, **speaking rate** is significant
  - $p = 0.0028$
- Thus the method still gains us insight into the feature
- From mixed-effect logistic regression with random intercepts for speaker
  - $N = 808$

# Implications

- Sequential temporal factors may indeed provide useful insights into language variation at various levels of linguistic structure
  - The fact that different temporal predictors emerge as significant for different variables makes available a number of new questions for study
    - What does it mean for some variable patterns like (ing) use to relate to pause structure/hesitancy and others (like copula/aux absence) to relate to speaking rate?
    - Further, do these differences provide insight into differential organization of these variables in speakers' grammars?
  - In sum, the treatment of sociolinguistic data in new ways (here, through a visualization technique) can lead to new insights

## In closing...

- SLAAP represents our attempts to ensure that individual collections of sociolinguistic data remain accessible and useful over time, and to reconsider the nature of our data
- So far, we hope we have been successful in articulating a proof-of-concept
- But there is very much more to do...

# Thank you

SLAAP: <http://ncslaap.lib.ncsu.edu/>

For a broader survey of SLAAP's features and architecture  
see the informal and not altogether complete user guide at:

<http://ncslaap.lib.ncsu.edu/userguide/>

Tyler: [tsk@uoregon.edu](mailto:tsk@uoregon.edu)