

The Sociolinguistic Archive and Analysis Project User Guide

Version 0.96 (second DRAFT – June 2009)

(minor updates since first draft)

Tyler Kendall, tsk3@duke.edu

0.	Why this document?	1
1.	What is the Sociolinguistic Archive and Analysis Project?	1
2.	User access and resource management.....	2
2.1.	User accounts.....	2
2.1.1.	NCLLP users	3
2.1.2.	Non-NCLLP users.....	3
2.2.	Resource management	3
2.2.1.	User access to archived materials.....	3
2.2.2.	User access to software features	4
3.	SLAAP's features	4
3.1.	Basic features.....	5
3.1.1.	Main library view	5
3.1.2.	Full record view.....	9
3.1.3.	Audio listen (and annotate).....	11
3.1.4.	Audio extraction and downloading	12
3.2.	Transcript features.....	13
3.2.1.	Browsing (for) transcripts	14
3.2.2.	Transcript viewing.....	15
3.2.3.	Transcript line analysis and phonetic analysis tools	17
3.2.4.	Transcript statistics and (semi-)auto-summarizing	18
3.2.5.	Exporting SLAAP transcripts	20
3.2.6.	Adding transcripts to SLAAP	21
3.3.	User features	21
3.3.1.	User preferences	21
3.3.2.	User forum	22
3.4.	Advanced and experimental features	23
3.4.1.	Search.....	23
3.4.2.	Variable extraction and coding (tabulation)	26
3.4.3.	Transcript-based speaker analysis tools.....	31
3.4.4.	Transcript-based lexical analysis tools.....	32
4.	Digitizing audio cassettes and adding data to SLAAP	34
5.	Transcribing	34
5.1.	Transcribing in Praat.....	35
5.1.1.	Praat, the TextGrid, and TextTiers.....	35
5.1.2.	Delimiting the utterances	38
5.1.3.	Orthographic conventions.....	38
5.1.4.	Recommended Praat settings.....	41
5.2.	Adding Praat transcripts to SLAAP	42
5.2.1.	Steps for adding Praat transcripts to SLAAP	42
6.	Browser and software requirements	46
7.	References and further reading	47
8.	Acknowledgements.....	48

0. Why this document?

This document is a working draft. Someday it will be a fully fleshed out user manual for the Sociolinguistic Archive and Analysis Project (SLAAP). Currently, it's mostly an overview of the SLAAP software. It aims to provide NCLLP members and other SLAAP users some help in navigating and making use of SLAAP's many features. While it surely fails along these lines in numerous places, it, hopefully, at least gives a sense of SLAAP's scope and intentions. For specific help, or if you have comments on SLAAP or this document, please contact the authors.

1. What is the Sociolinguistic Archive and Analysis Project?

At its most basic level, the Sociolinguistic Archive and Analysis Project (SLAAP; <http://ncslaap.lib.ncsu.edu/>) is a web-based archive for sociolinguistically relevant audio (and someday soon video) recordings. It is an initiative of the North Carolina Language and Life Project (NCLLP; <http://www.ncsu.edu/linguistics/ncllp/>) in partnership with the North Carolina State University Libraries (<http://www.lib.ncsu.edu>). As such, the recordings housed in the SLAAP archive are predominately sociolinguistic interview recordings conducted by the NCLLP. The largest group of users is the members and alumni of the NCLLP. That said, SLAAP also houses recordings conducted by other researchers and research groups, as well as other audio corpora for which SLAAP users have licenses or permissions. Many SLAAP users are researchers from other institutions who use SLAAP to access their own materials or to instantly share (where permitted) resources. If you are interested in having your data housed in SLAAP, or hope to access some of the current archive, please contact Tyler Kendall (tsk3@duke.edu) or Walt Wolfram (walt_wolfram@ncsu.edu).

At the same time as SLAAP is simply a web-based archive for speech data, it also represents an attempt to approach the storage, management, and analysis of language recordings in fundamentally new ways, leveraging web- and computer-based methods to enhance sociolinguistic practice. The theoretical and historiographical underpinnings of the project are spelled out elsewhere (cf. Kendall and French 2006, Kendall 2007a, Kendall 2008a, Kendall 2009) and not addressed in this document. Instead this document is intended as a *user manual*,

highlighting the various features of the archive software for its users. Section 2 briefly discusses how to gain access to the archive and how user accounts work in SLAAP. Section 3, the bulk of this document, overviews each of the major features in SLAAP. It is organized in parts, discussing “basic” features, “transcription” features, “user” features, and “advanced/experimental” features. Section 4 is currently a placeholder for a section on digitization and how users – NCLLP members in particular – add data to the archive; in the meantime, materials are available in the NCLLP’s Linguistics Lab (204 Tompkins Hall, NCSU) that outline the digitization and data entry process. Section 5 provides an outline and recommendations for building time-aligned transcripts using the open-source software Praat (Boersma and Weenink 2007; <http://www.fon.hum.uva.nl/praat/>) and describes how to add Praat-based transcripts to the SLAAP system. It also describes the transcript conventions used for SLAAP, though more extensive documentation is still needed.

Both this document and SLAAP, itself, are works-in-progress. Please note that screenshots may be inaccurate and that not all the features/pages of SLAAP are discussed in this document. At present, this user guide simply highlights the features and recommended use of the software. It is not (yet) a fully developed “help” guide or system. Users are urged to contact the author – or to post messages in the user forum (see section 3.3.2) – about software bugs, problems with data in the archive, ideas about new features, etc. Users are also invited to contribute to future versions of this manual (pretty please).

2. User access and resource management

2.1. User accounts

Access to the SLAAP software and archive is restricted to registered users. In order to use SLAAP you must have your own user id and password or a user id/password assigned to a class at an affiliated university, such as NCSU or Duke. (Note that a handful of tools based on SLAAP features are available as stand-alone, public resources from the <http://ncslaap.lib.ncsu.edu/tools/> web page.)

2.1.1. NCLLP users

All members of the NCLLP are able to have at least nominal access (see section 2.2 below) to the archive. To sign up for an account, use the “Request a SLAAP account” link on the NCLLP’s *stafftools* web page.

2.1.2. Non-NCLLP users

Sociolinguists and other researchers who are not affiliated with the NCLLP may still be allowed access to the archive under a number of different circumstances. Accounts are often made for outside researchers in order to share (when permissible) research materials, such as audio recordings, transcripts, or variable data. To an increasing degree, other researchers may store their recordings in SLAAP and will, of course, be given access to the SLAAP archive in order to interact with their own data.

2.2. Resource management

Since there are a number of ethical and copyright related issues to the storage and sharing of social scientific, human subject data, SLAAP has a fine-grained group-based resource management permissions system.

2.2.1. User access to archived materials

All SLAAP user accounts belong to one or more “user groups” in the SLAAP system. SLAAP controls access to materials on a per-project, per-group basis, where a “project” represents a collection of recordings and their associated data and meta-data (cf. Kendall 2008) organized around a particular field site or research project. So, for example, NCLLP users automatically belong to the “ncllp” group, which provides them access to all of the materials in SLAAP belonging to the NCLLP. A non-NCLLP researcher, who is given access to the Ocracoke recordings, for example, may belong to a group called “ocracoke” with access limited

to only those Ocracoke recordings. Some collections of recordings in SLAAP (such as the “NC Speech Samples” project) do not have copyright or human subjects constraints on them and, typically, are gratuitously made available to all SLAAP users.

In sum, SLAAP’s access control system allows for the archive to maintain compliance with copyright, licensing, and human subject constraints on the materials despite the wide variety of users and materials in the collection.

2.2.2. User access to software features

In addition to controlling users’ access to particular materials, the SLAAP access control system limits the level of access by sets of features. All user accounts can listen to and download those materials to which they have permissions, as well as read any associated transcripts, “public” notes (see section 3.1.3 on the annotation feature), and additional materials, such as uploaded Word documents, vowel plots and so on. Since the data in SLAAP are “live” – that is, meta-data about each recording, transcripts, and so on are editable – only certain user accounts have access to features like the sound file management pages, transcription uploading and editing, and variable tabulation and analysis. Many of the “advanced/experimental” features have very limited access, as they are not (yet) deemed useful to general users.

This is all to say, that your user account may not have access to many of the features outlined in this document (and that you may have access to features not discussed in this document). If you are interested in using some of the features outlined in this document that you don’t have access to, please let Tyler know.

3. SLAAP’s features

The presentation of SLAAP’s features is organized into a number of sub-sections:

- 3.1. *basic features*, such as the various “library” views of the data, and audio listening and downloading.

- 3.2. *transcript features*, having to do with the various presentation and interface formats to SLAAP’s transcripts, including the transcript-based, phonetic analysis features.
- 3.3. *user features*, such as the user preferences page, and the putative user forum.
- 3.4. *advanced/experimental features*, including the search feature, the variable extraction and coding (i.e. tabulation) features, and transcript-based speaker analysis tools for features, such as pause and speech rate.

In the following sections, these features are discussed and screenshots of actual web pages are given. Since different user accounts have access to different levels and types of data, and since the software is constantly being improved, these screenshots may not exactly match what your account experiences.

3.1. Basic features

SLAAP’s “basic” features center around users’ access to the core data of the SLAAP archive, the audio files, and data on their associated projects, speakers, and interview events.

3.1.1. Main library view

The main library view is the central web page of the SLAAP system. It is the page that you arrive at after logging in to SLAAP. A basic example of this page, for user *tsk3*, is shown in Figure 3.1.1.1, showing the first seven interview records for the Ocracoke project.

Through the main library page you can access almost all of SLAAP’s features. The very top of the page has various general purpose links and information, including links to the user forum (see 3.3.2), to your account settings (see 3.3.1), and to log out of the system (note that for security purposes users are automatically logged out 30 minutes after the web server last receives activity). This top part of the page – the “header” – is repeated on all of SLAAP’s pages but won’t be discussed further. Immediately above the main, gray-colored part of the page are links to any non-basic features that your account has access to (the user *tsk3* has access to variable

tabulations so, in Fig. 3.1.1.1, has a link for the tabulation summary features, see section 3.4.2). The absence of links here indicates that you do not have access to these additional features.

NC STATE UNIVERSITY [[User Forum](#)] [[tsk3](#) : [Acct](#) | [Logout](#)]
 [[Linguistics](#) | [Libraries](#) | [SLAAP Home](#) | [NCLLP Staff Tools](#)] (autologout if no activity at 1:39:26)
SLAAP v. 0.95 - Main Library Access [[@library.php](#)]
 Disk Usage: 0 kb

[[Tabulation Summary](#)]

NC SLAAP Archive: [Browse](#) | [Filter](#) | [Projects](#) | [Speakers](#) | [Transcripts](#) [[All Archive Search](#) | 1031 total records in SLAAP]
 [View: Long | Short] [Show Project Info?] [20 records at a time]
 [Project: Ocracoke | Showing 20 (of 67 records)] [Page: 1, 2, 3, 4 | >>]

Interview	Project	Speaker(s)	Interview Info	Media	Transcripts
ocr001 [Full Record]	Ocracoke	RWS White Male, Born 1920 Locality: Ocracoke, NC	Date: 02/25/1993 Interviewer(s): PA, KAH Language(s): English Contains: Sociolinguistic interview	ocr0010a [Listen Download] ocr0010b [Listen Download]	
ocr002 [Full Record]	Ocracoke	MB White Male, Born 1927 Locality: Ocracoke, NC	Date: 02/26/1993 Interviewer(s): PA Language(s): English Contains: Sociolinguistic interview	ocr0020a [Listen Download] ocr0020b [Listen Download]	
ocr003 [Full Record]	Ocracoke	JTO White Male, Born 1924 Locality: Ocracoke, NC	Date: 02/26/1993 Interviewer(s): PA Language(s): English Contains: Sociolinguistic interview	ocr0031a [Listen Download] ocr0031b [Listen Download] ocr0032a [Listen Download]	
ocr004 [Full Record]	Ocracoke	BOG White Male, Born 1915 Locality: Ocracoke	Date: 02/27/1993 Interviewer(s): WW, CC Language(s): English Contains: Sociolinguistic interview	ocr0040a [Listen Download] ocr0040b [Listen Download]	
ocr005 [Full Record]	Ocracoke	CW White Male, Born 1913 Locality: Ocracoke, NC	Date: 02/27/1993 Interviewer(s): PA Language(s): English Contains: Sociolinguistic interview	ocr0050a [Listen Download] ocr0050b [Listen Download]	
ocr006 [Full Record]	Ocracoke	EH White Female, Born 1911 Locality: Ocracoke, NC	Date: 02/24/1993 Interviewer(s): NSE, PA, MW, BR Language(s): English Contains: Sociolinguistic interview	ocr0060a [Listen Download] ocr0060b [Listen Download]	
ocr007 [Full Record]	Ocracoke	EH White Female, Born 1910 Locality: Ocracoke	Date: 12/14/1993 Interviewer(s): WW, MW Language(s): English Contains: Sociolinguistic interview	ocr0070a [Listen Download] ocr0070b [Listen Download]	
ocr008 [Full Record]	Ocracoke	BOS White Female, Born 1920	Date: 02/27/1993 Interviewer(s): NSE, MW	ocr0080a [Listen Download] ocr0080b [Listen Download]	

Figure 3.1.1.1 Main library view

The main use of this library view is intended to act like an online library card catalogue system for the interviews in the SLAAP archive. Each record (a row in the main table) shows the name of the interview, the project to which it belongs, and then information about the speaker(s) in the interview, and other information about the interview. Links are available for each record to a number of basic features, including the “full record view” (see 3.1.2), and the “listening” (section 3.1.3) and “downloading” (section 3.1.4) pages for each of the interview’s media files. If there were, links to these too would be shown here, under the transcripts column. For the interviews shown in Figure 3.1.1.1, there are no associated SLAAP transcripts.

A primary way users access the archived materials is by browsing through the materials via the library page. The pop up menu (in Fig. 3.1.1.1 showing “Ocracoke”) allows users to quickly select which set of resources they wish to browse. For user *tsk3*, SLAAP shows 20 records per page, so pagination links are available to the right of the pop up menu for paging through the available records. Users can also “filter” the available records by demographic information about the speakers. Figure 3.1.1.2, for example, shows the Ocracoke interviews with African American speakers.

The screenshot shows the SLAAP library interface with the following elements:

- Header:** NC STATE UNIVERSITY, [Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools], [User Forum] [tsk3 : Acct | Logout] (autologout if no activity at 1:50:06), SLAAP v. 0.95 - Main Library Access, [@library.php]
- Navigation:** [Tabulation Summary], [All Archive Search] | 1031 total records in SLAAP
- Filters:** [View: Long | Short] [Show Project Info?] [20 records at a time] [Project: Ocracoke] | Showing 5 (of 5 records) [Page: 1]
- Search Section:** Search for Interviews with a Specific Speaker: [Select the Speaker], Or, for Speakers Matching the Following Demographic Information: Ethnicity: Black | Sex: [Specific Sex] | YOB/Age: << Year of Birth << [Search]
- Table:**

Interview	Project	Speaker(s)	Interview Info	Media	Transcripts
ocr041_1 [Full Record]	Ocracoke	MB Black Female, Born 1904 Locality: Ocracoke, NC	Date: 03/16/1995 Interviewer(s): WW, KAH Language(s): English Contains: Sociolinguistic interview	ocr0411a [Listen Download] ocr0411b [Listen Download]	
ocr041_2 [Full Record]	Ocracoke	MB Black Female, Born 1904 Locality: Ocracoke, NC AB Black Female, Born 1915 Locality: Ocracoke, NC	Date: 07/20/1996 Interviewer(s): NSE, JJS Language(s): English Contains: Sociolinguistic interview	ocr0412a [Listen Download] ocr0412b [Listen Download]	
ocr041_3 [Full Record]	Ocracoke	MB Black Female, Born 1904 Locality: Ocracoke, NC AB Black Female, Born 1915 Locality: Ocracoke, NC	Date: 11/18/1996 Interviewer(s): NSE, EWG Language(s): English Contains: Sociolinguistic interview	ocr0413a [Listen Download] ocr0413b [Listen Download]	
ocr041_4 [Full Record]	Ocracoke	MB Black Female, Born 1904 Locality: Ocracoke, NC AB Black Female, Born 1915 Locality: Ocracoke, NC	Date: 03/11/1998 Interviewer(s): EWG Language(s): English Contains: Sociolinguistic interview	ocr0414a [Listen Download]	
ocr056 [Full Record]	Ocracoke	JB Black Male, Born 1936	Date: 07/25/1996 Interviewer(s): Language(s): English Contains: Sociolinguistic interview...	ocr0560a [Listen Download]	
- Footer:** [Page: 1]

Figure 3.1.1.2 Library, filter view

In Fig. 3.1.1.1, *tsk3*'s access to the Ocracoke project, 67 total records are available, while 5 – the only five with African American subjects – are available under the “filtered” view. **Note** that it is not recommended that you filter across all projects. Put differently, when used on the

entire archive (i.e. when the Project menu shows “[All Projects]”), the “filter” option can be extremely slow to load.

The “projects”, “speakers”, and “transcripts” links at the top of the library table bring you to alternative ways to browse the archive. An extracted screenshot from the projects page, centered on the Ocracoke project, is illustrated in Figure 3.1.1.3. Note that users with appropriate access privileges can associate (and even upload) publications and additional files to their relevant project. The speakers page is not illustrated here (and is less useful for browsing the archive). The transcripts page is discussed in section 3.2.1.

Access: [071807_Ling_Proming_Thomas.pdf](#) [NCSLAAP Access Only]

Thomas, Erik R. (2005). "Cues Used for Distinguishing African American and European American Voices". *Journal of the Acoustical Society of America*, 117.: 2458

Thomas, Erik R., and Jeffrey Reaser (Fc.), "An Experiment on Cues Used for Identification of Voices as African American or European American". *Michael D. Picone and Catherine Evans Davies (eds.), Language Variety in the South III*. Tuscaloosa, AL: University of Alabama Press
 Access: [071807_LAVIS_ms_30_Thomas & ReaserText_proposed_revision.doc](#) [NCSLAAP Access Only]

Thomas, Erik R., Norman J. Lass, and Jeannine Carpenter (Fc.), "Identification of African American Speech". *Dennis R. Preston and Nancy Niedzielski (eds.), Reader in Sociophonetics*. Cambridge, UK: Cambridge University Press
 Access: [071807_CambridgeAlAmlid.doc](#) [NCSLAAP Access Only]

Associated Files:

File	Size	Type	Notes	Owner
071807_GrantReading.doc	46 kb	Reading Passages, Wordlist		ethomas

Ocracoke Data-/Media- Entry Incomplete [[Show Full Info](#)]

Location: Ocracoke (Hyde County), NC USA **P.I.:** Walt Wolfram

Number of Speakers: 78 **Number of Interviews:** 67

URL: <http://www.ncsu.edu/linguistics/ncllp/sites/ocracokeisland.php>

Notes/Information:

Associated Publications:

Hazen, Kirk Allen (1994). *Subject-verb concord in post-insular vernacular varieties of English*. M.A. thesis, North Carolina State University

Schilling-Estes, Natalie (1996). *The Linguistic and Sociolinguistic Status of /ay/ in Outer Banks English*. Ph.D. dissertation, University of North Carolina at Chapel Hill

Schilling-Estes, Natalie, and Walt Wolfram (1999). "Alternative models of dialect death: Dissipation versus concentration". *Language*, 75.: 488-521

Vadnais, Janelle Chaundre (2006). "A cross regional study of locative to in North Carolina". M.A. thesis, North Carolina State University

Wolfram, Walt, and Natalie Schilling-Estes (1995). "Moribund dialects and the endangerment canon: The case of the Ocracoke Brogue". *Language*, 71.: 696-721

Wolfram, Walt, and Natalie Schilling-Estes (1996). "Dialect change and maintenance in a post-insular island community". *Focus on the USA*, ed. by Edgar W. Schneider. *Varieties of English around the World 16*. Amsterdam/Philadelphia: John Benjamins: 103-48

Wolfram, Walt, Kirk Hazen, and Jennifer Ruff Tamburo (1997). "Isolation within isolation: A solitary century of African-American Vernacular English". *Journal of Sociolinguistics*, 1.: 7-38

Wolfram, Walt, Kirk Hazen, and Natalie Schilling-Estes (1999). *Dialect Change and Maintenance on the Outer Banks. Publication of the American Dialect Society 81*. Tuscaloosa, AL/London: University of Alabama Press

Wolfram, Walt, Natalie Schilling-Estes, Kirk Hazen, & Chris (1997). "The sociolinguistic complexity of quasi-isolated Southern coastal communities". *Language Variety in the South Revisited*, ed. by Cynthia Bemstein, Thomas Nunnally, and Robin Sabino. Tuscaloosa/London: University of Alabama Press: 173-87

Associated Files:
None.

Ocracoke II Data-/Media- Entry Incomplete [[Show Full Info](#)]

Location: Ocracoke (Hyde County), NC US **P.I.:** Walt Wolfram

Number of Speakers: 51 **Number of Interviews:** 42

URL: <http://www.ncsu.edu/linguistics/ncllp/sites/ocracokeisland.php>

Notes/Information:

Associated Publications:
None.

Associated Files:
None.

Ohio DARE Re-survey Data-/Media- Entry Incomplete [[Show Full Info](#)]

Location: various (various County), OH US **P.I.:** Erik R. Thomas

Figure 3.1.1.3 A section of the “projects” page, centered on the Ocracoke project information

3.1.2. Full record and speaker record views

The full record page continues SLAAP’s imitation of an online library card system by showing a more complete record of the stored information for a given interview. Figure 3.1.2.1 shows the top part of the full record view for interview *prv007a*, an interview from the Princeville project. Notice that there are two media files for this interview, *prv007aa* and *prv007ab*; only the first is shown in the screenshot, the second is cut off from the bottom of the screen.

SLAAP v. 0.95 - Full Record Access - prv007a [@full_record.php]
 Disk Usage: 0 kb [Library]

prv007a [Go to: prv006 << * >> prv007b | Browse Project]

Project:	Princeville Princeville, Edgecomb County, NC, US Website: http://www.ncsu.edu/linguistics/nclp/sites/princeville.php
Speaker Info:	PEO black female, born 1964 (39 at time of interview) [Speaker Record]
Interviewer(s):	Rowe, Ryan (RR) Grimes, Andrew Gelvin Burley "Drew" (DG)
Interview Date:	09/26/2003
Media File(s):	prv007aa, prv007ab
Language(s):	English
Formats:	sociolinguistic interview
Interview Notes:	at Mayor's office

Associated Files:

File	Size	Type	Associated With	Notes	Owner
020608_PRV007atr.doc	44 kb	Transcript	prv007aa	Transcript of the beginning of the interview.	ethomas

prv007aa [Listen Download/Extract]	Data Overview
Audio Length: 46.02 min (~2761 sec) <small>Recalculate from Praat</small>	
Signal-To-Noise Ratio: Approx. 38 dB *okay quality* <small>(i.e. Approx. Audio Quality)</small>	
Transcript(s): Time range (in secs) 840-1430 Time range (in secs) 1980-2090	
Variable Tabulations: [Tab Variables] 3rdsgs - Needs started copula - In progress r - no status entry	
Digitization Metadata: File Specs: 44.1 khz 16 bit mono Creator: Amanda French Custodian: NCSU Libraries Digitization Date: 11/28/2005 Source Medium: Cassette	

Figure 3.1.2.1 A part of the full record page for Princeville interview prv007a

The full record view gives more complete information about the interview than shown in its record on the main library screen (section 3.1.1). Additionally, you can access (and with

proper permissions upload) associated files on this page – anything from Word document transcripts, notes, interview protocols and reading passages, photographs, etc.

For each media file, the full record view displays an array of information, from automatically generated information about the audio (such as its length and an approximation of its signal-to-noise ratio) to metadata about the original recording and the digitization process. Links are also provided to any transcripts and variable tabulation sheets that are in the system (the user, *tsk3*, has access to SLAAP’s tabulation features, or else wouldn’t have links to tabs, see 2.2.2). There are also automatically generated “data overview” time lines for each of the media files, showing the temporal locations of any transcripts and analysts “annotations” (see 3.1.3. on annotations). These are clickable, and bring you to the transcript or the location in the audio of a given annotation.

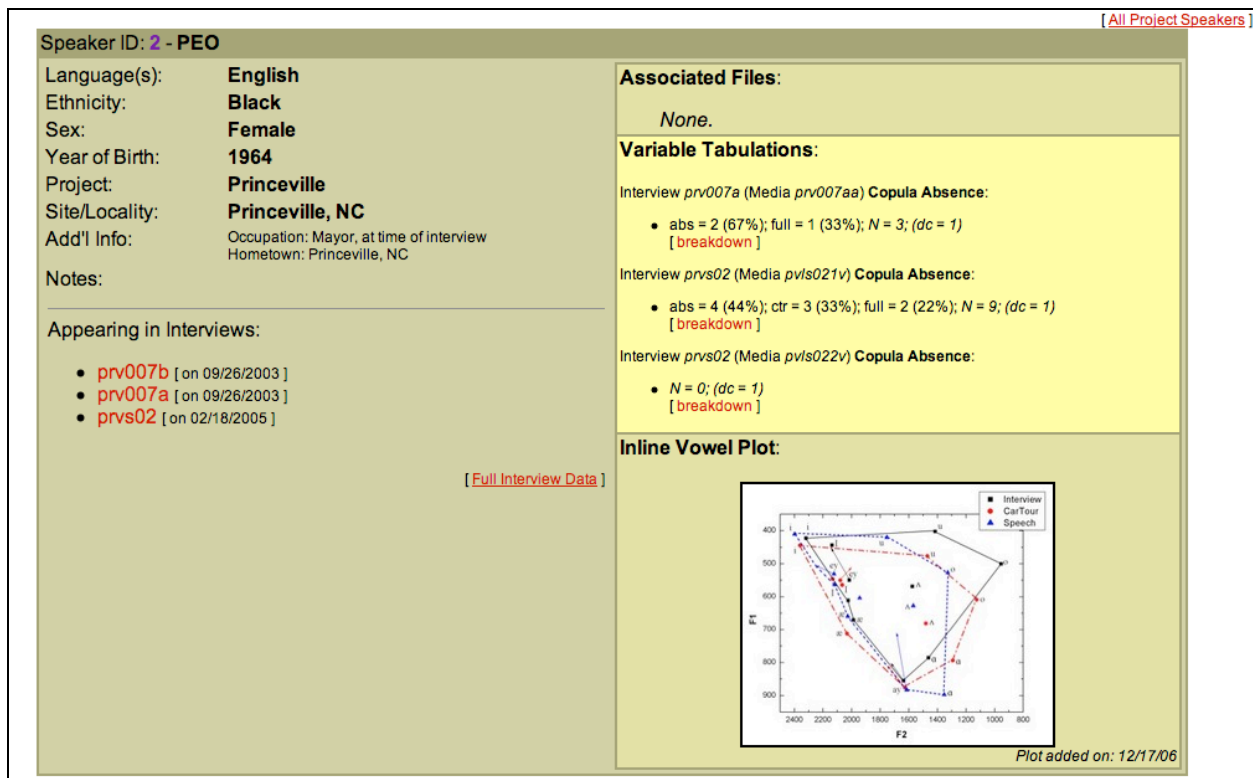


Figure 3.1.2.2 Speaker record for a speaker from Princeville

Similar to the full record view for interviews are the speaker record pages. These pages give complete information about each speaker’s recording in SLAAP, including links to the interviews in which he or she appears, summary variable tabulation information, if any exist (see

section 3.4.2 on SLAAP’s variable extraction and coding features), and additional uploadable information, such as a vowel plot. Figure 3.1.2.2 demonstrates this page for a speaker from Princeville. A speaker’s speaker record page can be reached from a number of pages within SLAAP, including from links on the full record page (this is visible in Fig. 3.1.2.1 above) and the speakers summary page (discussed, but not shown, in section 3.1.1).

3.1.3. Audio listen (and annotate)

The “audio listen” page is the primary interface to the audio files. This is shown in Figure 3.1.3.1. An mp3 version of the archive quality wav file is provided to the user using the browser’s audio player plug-in. All of SLAAP’s analysis and phonetic features (such as the generation of the spectrogram in Figure 3.1.3.1) always use the full quality, wav-format audio recordings. However, mp3 versions of the audio files are often presented to the user to enable much faster downloading. In order for SLAAP’s time-aligned connectivity to the audio player to work properly, your browser must use Apple’s QuickTime player as its plug-in (QuickTime is available for free from <http://www.apple.com/quicktime/download/> and installation should be relatively straightforward following the instructions).

As illustrated in Fig. 3.1.3.1, users can associate time-stamped notes, called “annotations” in SLAAP terminology, to any point in the audio. These notes are searchable (see section 3.4.1 on SLAAP’s search capability) and can be used to instantly return to the moment in the audio to which they point. (Note that you often have to wait for the entire audio file to load before the links will move the audio player’s cursor to the correct time.) Users can make private notes, which are only available to themselves, group-level notes, which are available to all the members of the specified user group (see section 2.2.1 on user groups), or “public” notes, which are available to all SLAAP users. To delete a note, you simply click the “X” link following the note text. Only the creator of a note can delete it.

Finally, users can, for lack of a better word, “zoom” into areas of the audio that require finer attention. It extracts a 1, 2, 5, or 10 second audio segment centered on the specified time. The “zoom” feature has two primary benefits. First, by extracting a short stretch of audio, it makes it much easier to repeatedly listen to the same stretch of talk. It also creates a spectrogram that may help in determining characteristics of the talk in question.

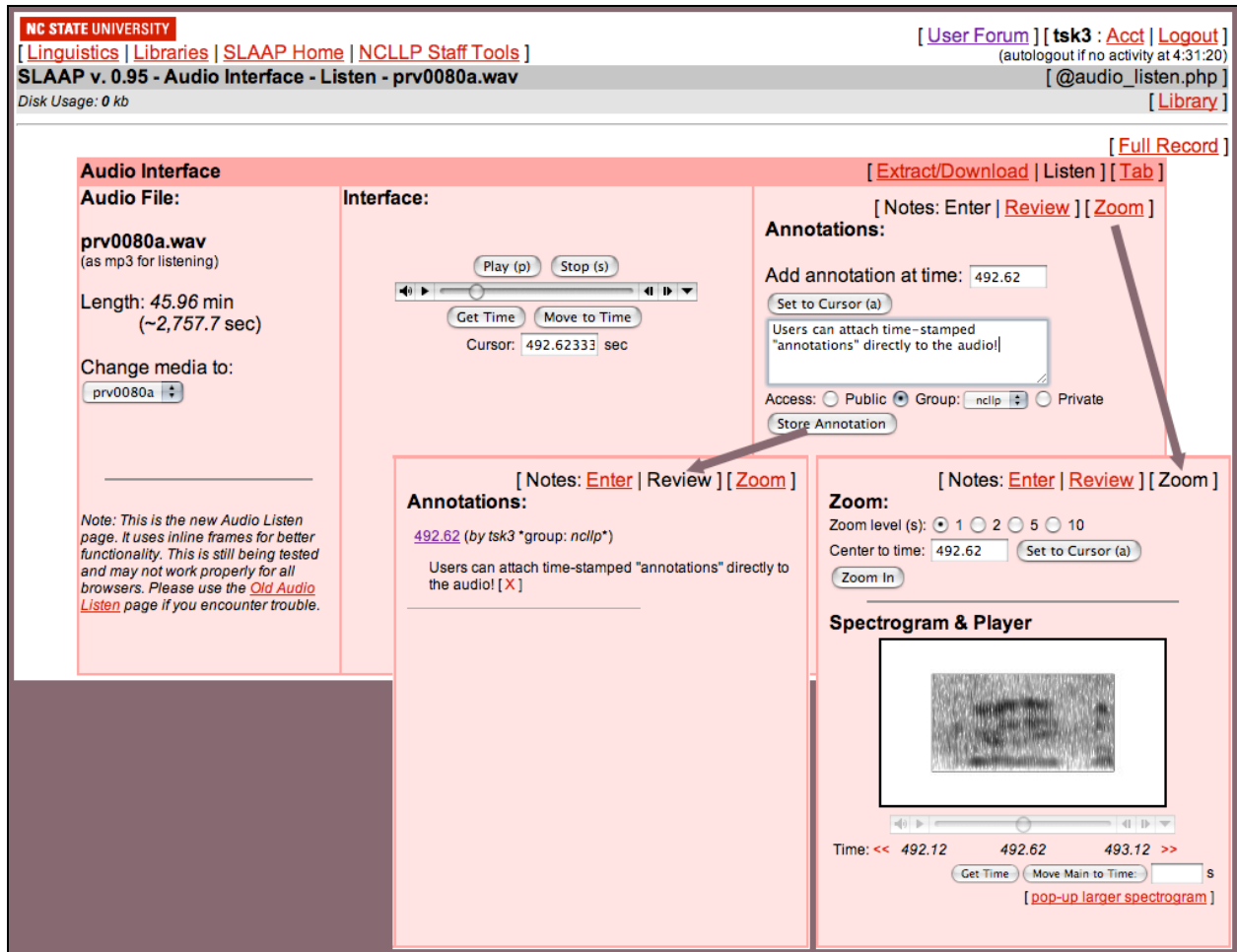


Figure 3.1.3.1 Audio listen page, demonstrating annotations and spectrogram generation

3.1.4. Audio extraction and downloading

Of course, much use of SLAAP centers on the downloading of segments of audio for analysis on the user’s local computer. The audio extract/download page, shown in Figure 3.1.4.1, allows you to export the archive quality wav files from the server to your local computer. This page is relatively self-explanatory.

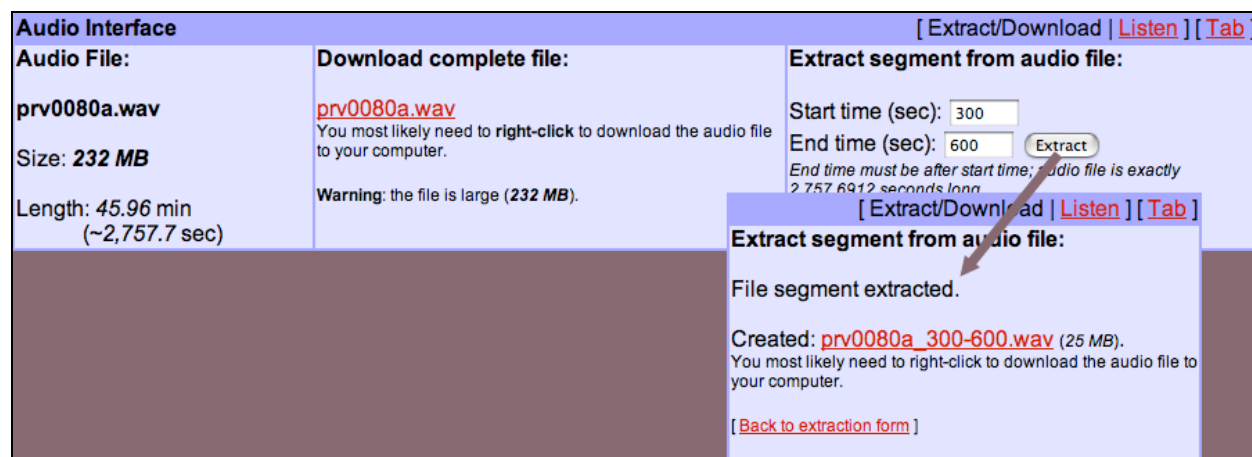


Figure 3.1.4.1 Audio download and extraction features

Note that all times are entered in seconds. **Note** also that when you extract a segment from an audio file, the internal time-stamping of the extracted segment is zeroed; it does not “remember” where in the larger audio file it was extracted from. That is, time-stamps obtained from the extract with, for example, Praat, will be specific to that extracted segment, with a start time of 0. The time range in terms of the original audio is recorded in the filename generated by SLAAP. This “zeroing” of the time domain is crucially important for time-aligned transcription or notes made on the excerpt. Within Praat there are ways to adjust the time domain of an audio file. SLAAP’s transcription upload feature allows you to offset the time domain of transcript for an extracted stretch of audio (see 5.2).

3.2. Transcript features

While only a very small proportion of the archive is transcribed, much of SLAAP’s full potential is realized through the transcription model. This model has been outlined elsewhere (see Kendall 2006-2007, 2007a, 2009) and is discussed in further detail in section 5 of this document, but the basic gist is that SLAAP transcripts are built using Praat in order to obtain fine-grained time-alignment. All speech is segmented such that a transcript “line” corresponds to uninterrupted phonation on the part of a single speaker (any silence longer than ~ 60 ms is segmented as its own “pause” line, again see Kendall 2006-2007, 2007a, 2009).

3.2.1. Browsing (for) transcripts

SLAAP’s transcripts are reachable from a number of different pages. In all cases, the transcripts are named with the following convention: the name of the media file, an underscore, the start time of the transcript (rounded to whole seconds), another underscore, and then the end time of the transcript (again rounded to whole seconds). Occasionally, transcript names will be shown prefaced with a “w_” (e.g., *w_prv0110a_1000_1600* for transcript *prv0110a_1000_1600*); this is an artifact of how these data are stored in the database (*w_prv0110a_1000_1600* and *prv0110a_1000_1600* are the exact same transcript).

prv011 [Full Record]	Princeville SK Black Male, Born 1948 Locality: Princeville, NC	interview... Date: 10/03/2003 Interviewer(s): RR, DG Language(s): English Contains: sociolinguistic interview, ?	prv0110a [Listen Download] prv0110b [Listen Download]	prv0110a_1000_1600
prv012	Princeville BP	Date: 10/25/2003	prv0120a [Listen Download]	

Figure 3.2.1.1 Location of transcript links from the main library view

[prv011](#) [Go to: [prv010](#) << * >> [prv012](#) | Browse Project | Edit Data]

Project:	Princeville Princeville, Edgecomb County, NC, US Website: http://www.ncsu.edu/linguistics/nclp/sites/princeville.php
Speaker Info:	SK black male, born 1948 (55 at time of interview) [Speaker Record]
Interviewer(s):	Rowe, Ryan (RR) Grimes, Andrew Gelvin Burley "Drew" (DG)
Interview Date:	10/03/2003
Media File(s):	prv0110a , prv0110b
Language(s):	English
Formats:	sociolinguistic interview, ?
Interview Notes:	
Associated Files:	None.

[+]

prv0110a [Listen Download/Extract]	Data Overview
Audio Length:	45.84 min (~2750 sec) <small>Recalculate from Praat</small>
Signal-To-Noise Ratio: <small>(i.e. Approx. Audio Quality)</small>	Approx. 39 dB *okay quality*
Transcript(s):	Time range (in secs) 1000-1600 [+]
Variable Tabulations:	[Tab Variables]
Digitization Metadata:	No variable tabulations for this media file File Specs: 44.1 khz 16 bit mono

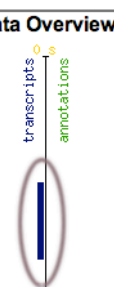


Figure 3.2.1.2 Locations of transcript links from the full record view

Figure 3.2.1.1 illustrates how transcripts can be found from the SLAAP’s main library view. Figure 3.2.1.2 illustrates where available transcripts are located on an interview’s full record page. When a transcript exists for a media file within a given interview, a link to it is available from both of those record views. SLAAP’s transcripts overview page is available for browsing all of the transcripts available (to your account) in the archive. The page is available from the “transcripts” link located at the top of the main library table on the main library page. This page, here showing transcripts from the Princeville project, is illustrated in Figure 3.2.1.3.

[[Search Transcripts](#) | Show Transcripts for: Princeville]

6 transcripts are currently available to you for Princeville, covering 1,600 seconds (about 0.44 hours) of audio (SLAAP has 128 total transcripts).

Transcript	Project	Interview/Media	Speaker(s)	Inter-viewer(s)	Length [sort]	Time Range (s) (% of Media File)	Num Lines	Special
prv007aa_840_1430	Princeville	prv007a / prv007aa	PEO	RR	590 sec (9.83 min)	840 - 1430 (21%)	663	summarize stats
prv007aa_1980_2090	Princeville	prv007a / prv007aa	PEO	RR	110 sec (1.84 min)	1980 - 2090 (4%)	170	summarize stats
prv0110a_1000_1600	Princeville	prv011 / prv0110a	SK	RR	600 sec (10 min)	1000 - 1600 (22%)	877	summarize stats
pvls015f_573_697	Princeville	prvs01 / pvls015f	SK	RJ	124 sec (2.07 min)	573 - 697 (10%)	131	summarize stats
pvls015f_840_884	Princeville	prvs01 / pvls015f	SK	RJ	44 sec (0.73 min)	840 - 884 (4%)	40	summarize stats
pvls021v_0_132	Princeville	prvs02 / pvls021v	PEO		132 sec (2.2 min)	0 - 132 (60%)	144	summarize stats

[sort by [added date](#)]

Figure 3.2.1.3 The transcripts overview page, showing transcripts available for Princeville

3.2.2. Transcript viewing

The basic transcript view, demonstrated in Figure 3.2.2.1, is formatted in a vertical, or “script-like”, format in fairly traditional ways (see, e.g., Ochs 1979). In a transcript’s initial view, each line is numbered, and is displayed with its start and end times in brackets. (See section 5 for transcript conventions.) Various check boxes are available to change the basic formatting of the transcript. You can, for example, hide line numbers or blank lines (which represent pauses). In Fig. 3.2.2.1, overlapping speech (indicated with square brackets) is indented. Other displays are available through the “Display” popup menu. Some of these are illustrated in Figure 3.2.2.2. Note that not all of the options work for all of the transcript displays.

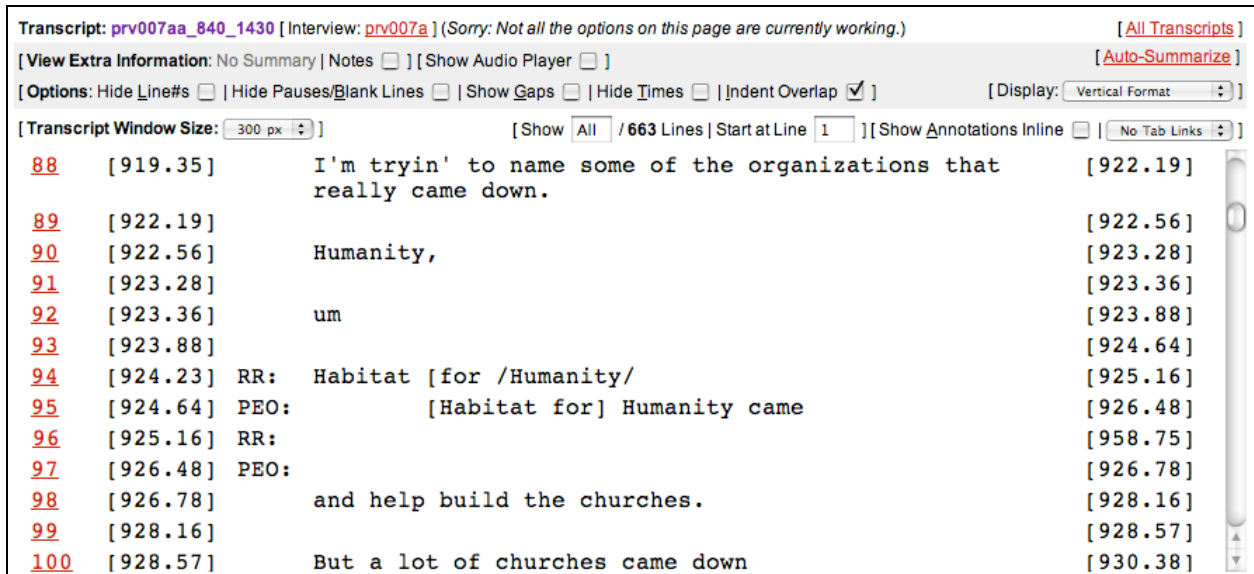


Figure 3.2.2.1 A basic transcript view

Some of the alternative transcript displays, in particular the graphical ones, are experimental. Discussions of them will be added to this document at a future date (in the meantime, see, e.g., Kendall 2007a, 2008b, 2009).

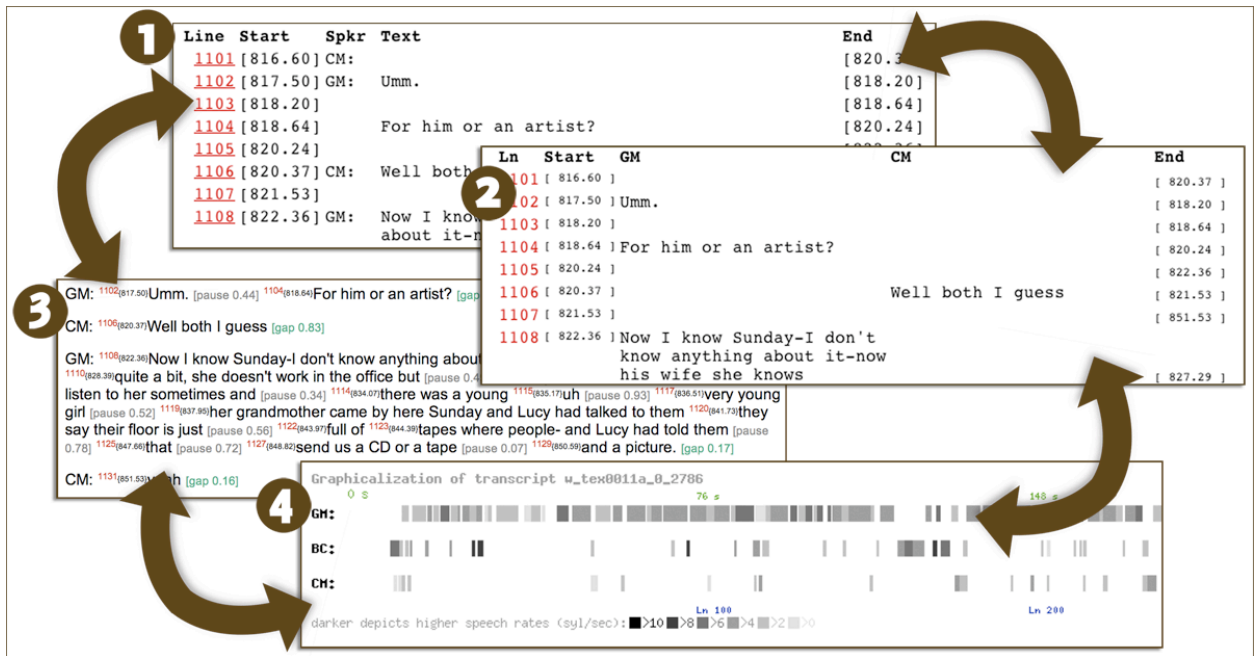


Figure 3.2.2.2 Various transcript views (from Kendall 2008a: 342, Fig. 2)

3.2.3. Transcript line analysis and phonetic analysis tools

The line numbers in the text-based transcript views are links that bring you to a line analysis page. There are a number of features available from this page. Some of these features are illustrated in Figure 3.2.3.1.

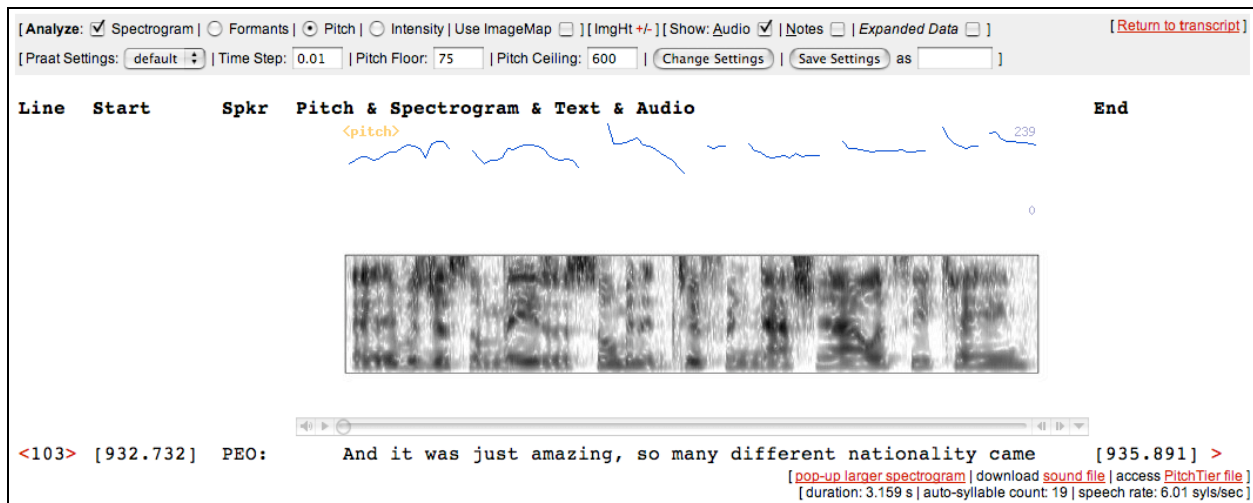


Figure 3.2.3.1 Line analysis view

Fig. 3.2.3.1 includes the audio for the line, a spectrogram and pitch track, both generated directly from the extracted audio. The settings used to generate the analysis, here on pitch, are shown under the heading “Praat Settings”. These settings are completely customizable and can be stored in your account for easy reuse (to delete stored settings, see section 3.3.1 on the user preferences page). You can page through the lines using the angle bracket links located at the line number and end time (these links skip blank lines so take you to the previous or next non-blank line). Not shown in Fig. 3.2.3.1 are the ImageMap, line Notes, and Expanded Data features. The ImageMap checkbox makes the pitch, intensity, or formant track clickable – allowing you to retrieve quantitative data directly by clicking on points in the plot (note that the formant tracking feature still needs some work). The “Notes” checkbox shows line-level notes in an editable text field. The “Expanded Data” checkbox brings up various experimental features (such as a part-of-speech tagging feature and an extremely experimental intonation transcription/analysis algorithm). You probably won’t find these “expanded” features useful without first talking to Tyler.

At the bottom right of the line information is more information about the line, including its length and speech rate in terms of syllables per second. SLAAP attempts to automatically count the syllables in each line's utterance (see Kendall 2007a), but can be over-ridden when wrong (through the "Expanded Data" checkbox). Also at the bottom right are links that allow you to download various pieces of data generated through the line analysis. You can also open a larger view of the spectrogram in its own browser window.

While not discussed further here (yet), this page also allows the comparison of multiple lines. At present, you must reach this page via a transcript search (see 3.4.1) to compare multiple lines.

3.2.4. Transcript statistics and (semi-)auto-summarizing

From both the transcripts overview page and each transcript page, you can access the "transcript summarizer" page. This page contains a number of related features, which allow SLAAP to help you explore the content of a given transcript. The "stats" feature – available from the "stats" link in the transcripts overview page or from the link at the top left of all the views on the summarizer page – provides data on all of the speakers' talk in the transcript, including the number of words spoken by each speaker and various metrics giving data on the length of talk by each speaker. The stats feature is illustrated in Figure 3.2.4.1. On the left-hand side of this page (shown in Fig. 3.2.4.1) are options available for all the features of this page.

Two different types of content "summaries" are available, through the "Summary Type" links on the left. The "High Freq." summary uses the selected N-gram level to pull out transcript lines with 2 or more of the most frequent N-grams (an N-gram is a sequence of N words, so a "bigram" is a two-word sequence, a "trigram" a three-word sequence). The "Int. Contr." (or "Interviewer Contributions") summary type attempts to give you a sense of the transcript's talk by displaying all of the interviewers' contributions that match certain criteria (currently, lines longer than 1 second contain a question word). The first (and currently only) SLAAP working paper (Kendall 2007c) discusses the "High Freq." summary method at some length.

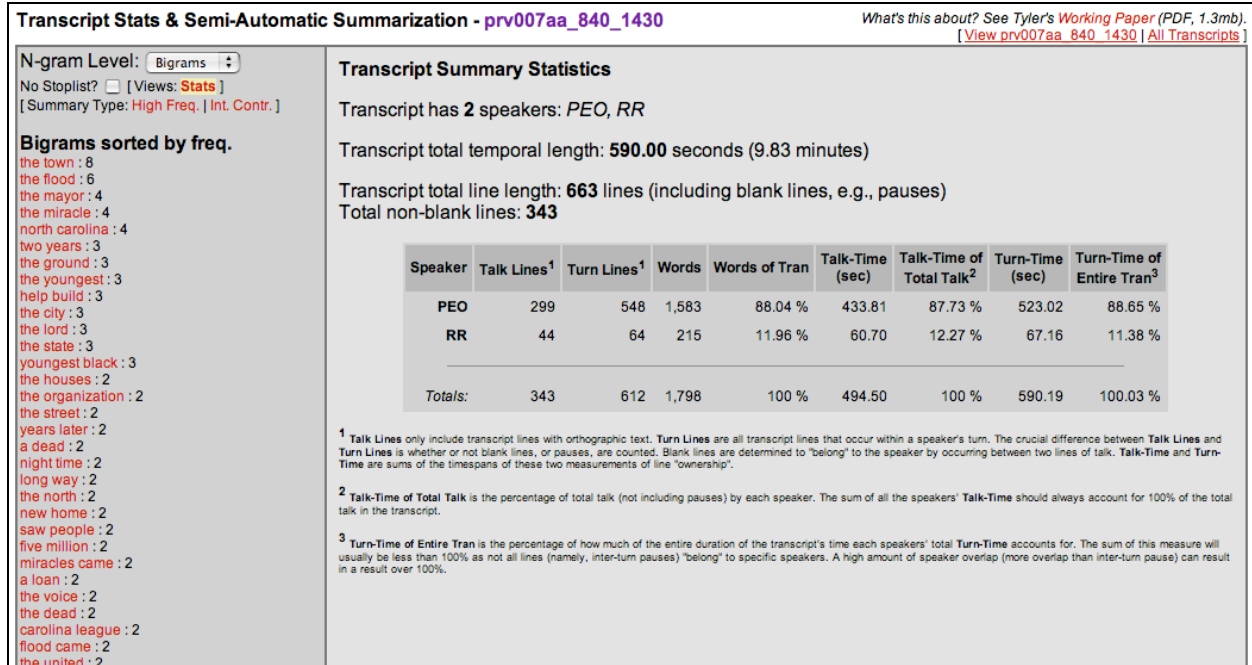


Figure 3.2.4.1 Transcript summary statistics and summarizer options

The frequency list of N-grams available on the left (again, see Fig. 3.2.4.1) are linked to a concordance feature. Clicking on one of these N-grams, such as “the mayor” generates a view as shown in Figure 3.2.4.2. A concordance is a method to provide keywords or matched phrases in context, and as illustrated in Fig. 3.2.4.2, the keyword or phrase is centered and highlighted within its matrix utterance. A time-line is also created showing where the phrase appears in the transcript, which is itself shown temporally situated within its media file.

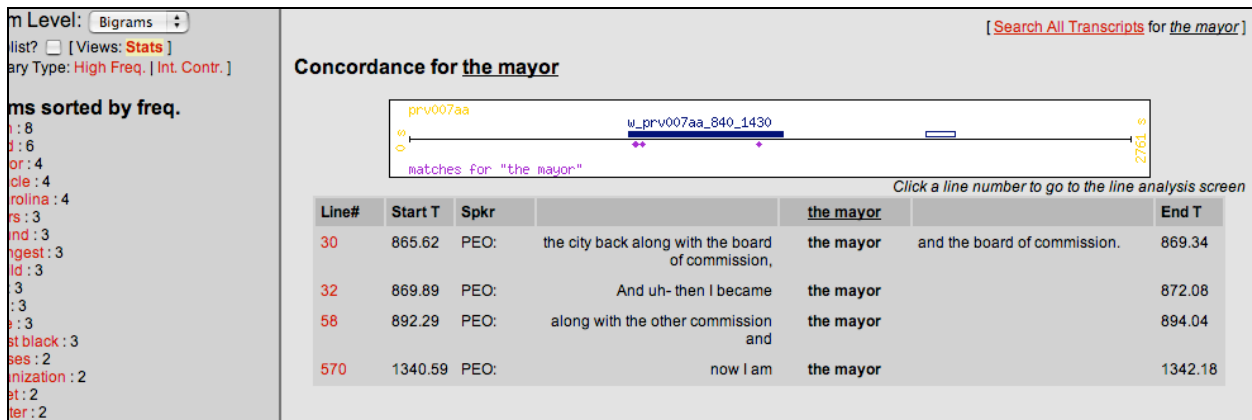


Figure 3.2.4.2 Concordance for “the mayor” from the summarizer page

Since the features on this page work only for the selected transcript, remember that the information generated here is only for the segment of talk covered by the transcript. The timeline in Fig. 3.2.4.2 shows, via the non-filled in rectangle, that there is a second, shorter transcript available for the current media file. However, the contents of this second transcript are not available here – if there were any instances of “the mayor” in that second transcript (there aren’t), they would not appear here.

3.2.5. Exporting SLAAP transcripts

Transcripts in SLAAP can be exported into Praat TextGrid format or simple tab-delimited text format by users with “manage transcript” access. If this option is available to you, a link will be available adjacent to the “Auto-Summarize” link at the top right of each transcript’s page (as illustrated in Figure 3.2.5.1) and from the entry for each transcript in the transcripts overview page (as illustrated in Figure 3.2.5.2).

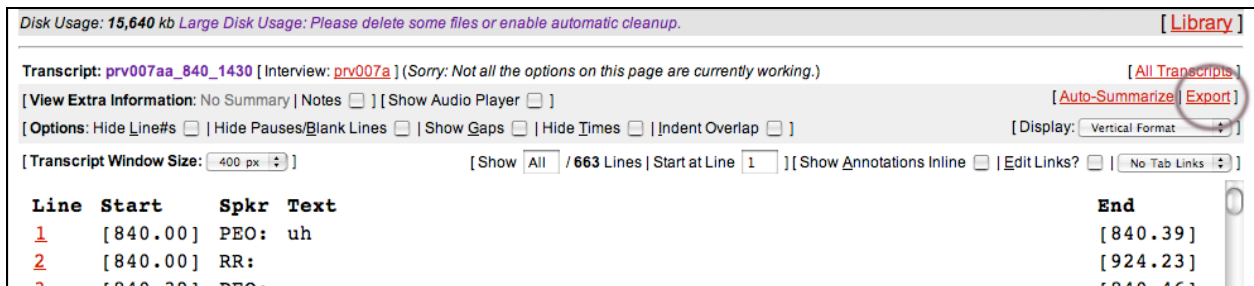


Figure 3.2.5.1 Location of the export link on a transcript’s page

prv0110a_1000_1600	Princeville	prv011 / prv0110a	SK	RR	(1.84 min) 600 sec (10 min)	1000 - 1600 (22%)	877	stats export summarize stats export
pvlis015f_573_697	Princeville	prvs01 / pvlis015f	SK	RJ	124 sec	573 - 697 (10%)	131	summarize

Figure 3.2.5.2 Location of the export links on the transcripts overview page

3.2.6. Adding transcripts to SLAAP

See section 5.2 in the general transcription section for instructions on how to add transcripts to the SLAAP archive. Note that only accounts with “manage transcript” access can upload transcripts.

3.3. *User features*

SLAAP has a number of features targeted primarily at the “user experience”. Those that don’t fit into the other categories of this guide are discussed here.

3.3.1. User preferences

Your user preferences page is available from the “Acct” link at the very top right of the header on every page in SLAAP. It is next to the “Logout” link. The user preferences page lets you customize certain aspects of your default SLAAP library view, such as how many records are displayed and whether your browsing view should default to a specific project. A screenshot from the user preferences page, for user *tsk3*, is shown below in Figure 3.3.1.1.

On your user preferences page, you can also view which user groups you belong to (see 2.2.1) and which projects you have access to, as well as what levels of software access you have (see 2.2.2). You can change your password by clicking the button.

As you navigate through SLAAP, the system generates a number of temporary files for you. These can be things like audio excerpts, spectrogram images, and numerical pitch data. These accumulate over the course of your SLAAP-use and generate the “Disk Usage” number that appears on the header of all of your pages. The default setting for SLAAP is that these files are automatically “cleaned up” (i.e., deleted) for you whenever you log out. Through the “File Management” section of this page, you can review and access (or delete) any of the files that exist for your account. You can also disable (or re-enable) the automatic clean up feature. If you disable automatic clean up, please be sure to manually delete your files periodically as they can accumulate rather quickly.

You are logged in as **tsk3**.

Main Settings		File Management
Full Name	Tyler kendall	My Files [Show]
Email	tsk3@duke.edu	[Auto Cleanup is on [Disable]]
Access Level(s)	read, tabulation	
Groups and Permissions	Group: ncllp Accessible Projects: Abaco Island; Asheville 1974; Beech Bottom; Bertie County; Cary; Durham Public Schools; Ex-Slave Recordings; Graham County; Harkers Island; Hickory; Hyde County; Johnstown, OH; Miscellaneous NC; Miscellaneous non-U.S.; Miscellaneous U.S.; NC Speech Samples; NSF-funded recordings; Ocracoke; Ocracoke II; Ohio DARE Re-survey; Pearsall, TX; Princeville; Raleigh; Roanoke Island; Robeson County; Roseburg, Oregon; Santa Barbara Corpus; Siler City; Texana; Türkçe ünüüler akustigi; UK Samples; Warren County; Western Reserve; Wilmington 1973; Wilson, NC; Zebulon Middle School	
Default Research Project	[All Projects in Library] ▾	
Default Library View	<input checked="" type="radio"/> Long (More Info) <input type="radio"/> Short (Less Info)	
Default Number of Records to Display in Library	20 [Update]	
Password	[Change Password]	

Stored Praat Settings

To save your own praat settings, use the form on the *analyze line* pages. The following are settings that you have stored. You may delete unwanted settings here.

Formant Settings	No formant settings stored.
Pitch Settings	No pitch settings stored.

Figure 3.3.1.1 SLAAP’s user preferences page for *tsk3*

Your user preferences page also has a section on your “Stored Praat Settings”. These are the settings that you may have stored on the transcript line analysis pages (see 3.2.3). You can review these here and delete any that you no longer need.

3.3.2. User forum

The SLAAP user forum is a relatively new (and unused) feature in SLAAP. It is intended to provide SLAAP users with a secure place to discuss various aspects of SLAAP and sociolinguistic practice in more general terms. The forum is organized around a number of categories and users can create or respond to threads within those categories. One screenshot of the forum, showing the current categories, is included here as Figure 3.3.2.1.

[[Start a New Thread](#)]

This is the new SLAAP user forum. It's a work in progress, but is intended to give everyone an easy and organized way to communicate about SLAAP issues and features, as well as general sociolinguistic and variationist practice... Please select a category.







Category	Last Post	Threads	Posts
 Announcements A forum for announcements and general (SLAAP-related) communications	Apr 12, 7:58 pm from Tyler Kendall New User Forum! »	1	1
 Suggestions & Bugs Help Tyler improve SLAAP. Give your feedback and report bugs.	May 29, 5:13 pm from Tyler Kendall Transcript Audio Playing »	4	7
 Help! Need help with a feature? Post about it here.	No messages yet	0	0
 Variable Tabbing Questions or ideas about Coding and Extracting Variables?	No messages yet	0	0
 Transcribing Questions or ideas about Transcribing?	Apr 17, 11:10 am from Leah White pronunciation »	1	1
 Misc. Miscellaneous.	Apr 20, 6:41 pm from Tyler Kendall Re: project site total hours »	1	2

Figure 3.3.2.1 The categories view of the SLAAP user forum

Tyler is happy to expand this feature as people desire. Right now, it is mostly intended as space for SLAAP users to comment on SLAAP related issues, such as bugs or feature requests.

3.4. *Advanced and experimental features*

In this section a variety of SLAAP's features are discussed. Some of these – such as search (section 3.4.1) and variable tabulation (section 3.4.2) – are central features for many SLAAP users. Others are less used and less useful. Sections 3.4.3 & 3.4.4 discuss features that Tyler has designed around his dissertation work (Kendall 2009) and for other experimental purposes. You're welcome to explore these features, with the caveats that they may not work entirely and that you should discuss these features with Tyler before using them for actual research.

3.4.1. Search

SLAAP has a comprehensive search page that will search a swatch of SLAAP's data for matching text. This page is accessible from the "All Archive Search" link at the top of the main library screen. It is also available from a number of links on other pages (such as from the summarizer's concordance page (see 3.2.4). Figure 3.4.1.1 illustrates the main search form.

Figure 3.4.1.1 SLAAP’s search form

The search feature is still slightly under construction. Please note the following. Tabulation data (see 3.4.2) are not yet included in the scope of a search. The “Limit search to” pop ups only limit searches within transcripts (sorry!). All searches are case-insensitive.

Interviews: Matches for Princeville [4 Results]

Interview	Project	Speaker(s)	Interview Info	Media	Transcripts	Notes
prv015 [Full Record]	Princeville	TVJ Black Female, Born 1982 Locality: Princeville, NC	Date: 03/06/2004 Interviewer(s): RR Language(s): <i>English</i> Contains: <i>Sociolinguistic interview</i>	prv0150a [Listen Download] prv0150b [Listen Download]		Possibly interviewed in her home in Princeville, North Carolina.
prv017 [Full Record]	Princeville	LB Black Female, Born 1984 Locality: Princeville, NC	Date: 03/06/2004 Interviewer(s): RR, KD Language(s): <i>English</i> Contains: <i>Sociolinguistic interview</i>	prv0170a [Listen Download] prv0170b [Listen Download]		Interviewed in Princeville, North Carolina.
prv030 [Full Record]	Princeville	TNB Black Female, Born 1986 Locality: Princeville, NC	Date: 06/11/2004 Interviewer(s): RR, KD Language(s): <i>English</i> Contains: <i>Sociolinguistic interview</i>	prv0300a [Listen Download] prv0300b [Listen Download]		Location of interview unclear, but apparently in Princeville, North Carolina.
prvs02 [Full Record]	Princeville	PEO Black Female, Born 1964 Locality: Princeville, NC	Date: 02/18/2005 Interviewer(s): DG Language(s): <i>English</i> Contains: <i>(political) speech</i>	pvls021v [Listen Download] pvls022v [Listen Download] pvls023v [Listen Download]	pvls021v_0_132	Mayor giving speech and introductions at the Princeville birthday celebration

[[Back to top](#)]

Speakers: Matches for Princeville [6 Results]

Spkr ID	Speaker	Project	In Interview(s)	Ethnicity	Gender	YOB age	Add'l Info	Notes
2	PEO	Princeville	prv007a,	black	female	1964	Occ: Mayor, at time of	

Figure 3.4.1.2 Some search results for “Princeville”

Figure 3.4.1.2 shows the beginning of the results for a search of “Princeville” through all the searchable data-types. Notice from Fig. 3.4.1.2 that the number of matches appear next to

the data-type in the “Search what?” area of the search form. There are 4 interview matches, 6 speaker matches, 29 instances in transcripts, 1 annotation match, and 0 matching forum messages among the data. Additionally, in the search result screen, the “Search What?” data-types are clickable and will scroll your browser window to the matches for that data-type. Note that the matching interview records are displayed in a similar fashion to their display in the main library page, but here notes are also included (the rightmost column) whereas in the main library view interview notes are not shown (though they appear in the full record view for each interview).

Transcript results, as shown in Figure 3.4.1.3, are displayed in a concordance format. You can click on a line number to go to that line in the transcript (note that these links do not take you to the line analysis page, but to the transcript itself). By checking the boxes in the leftmost “Sel” column you can analyze multiple lines in the transcript line analysis page by clicking the “analyze selected lines from this transcript” link at the bottom of each concordance. See section 3.2.3 on the line analysis page.

Transcripts: Concordance for Princeville

Results found in: [w_prv007aa_840_1430 \(7\)](#), [w_prv007aa_1980_2090 \(4\)](#), [w_prv0110a_1000_1600 \(8\)](#), [w_pvls015f_573_697 \(4\)](#), [w_pvls015f_840_884 \(1\)](#), [w_pvls021v_0_132 \(5\)](#)

Transcript: prv007aa_840_1430				Click a line number to go to the transcript at that line [7 Results]			
Sel	Line#	Start T	Spkr		Princeville		End T
<input type="checkbox"/>	25	860.17	PEO:	for the town of	Princeville	in two thousand.	862.02
<input type="checkbox"/>	41	879.03	PEO:	the town of	Princeville	after res-	880.75
<input type="checkbox"/>	113	946.62	PEO:	the bottom line was the goal was to rebuild	Princeville	back	950.54
<input type="checkbox"/>	117	952.94	PEO:	We have rebuilt	Princeville	back	954.67
<input type="checkbox"/>	119	954.95	PEO:	And I say ninety-eight percent of	Princeville		957.12
<input type="checkbox"/>	150	979.46	RR:	/	Princeville	/	979.97
<input type="checkbox"/>	572	1342.38	PEO:	of the town of	Princeville	.	1343.93
[analyze selected lines from this transcript]				Check the boxes and click the link to the left to analyze select lines from this transcript			

Transcript: prv007aa_1980_2090				Click a line number to go to the transcript at that line [4 Results]			
Sel	Line#	Start T	Spkr		Princeville		End T
<input type="checkbox"/>	5	1982.03	PEO:	Southern Terrace,	Princeville	, then Tarboro	1984.33
<input type="checkbox"/>	9	1985.16	PEO:	we just didn't have	Princeville	address	1987.30
<input type="checkbox"/>	13	1988.44	RR:	so, other than that /	Princeville		1989.91
<input type="checkbox"/>	16	1989.91	RR:		Princeville		1990.55

Figure 3.4.1.3 Transcript search results for “Princeville”

Annotation results, as shown in Figure 3.4.1.4, are displayed with the matching word(s) highlighted. See section 3.1.3 for more about the “annotation” type of notes. Clicking on the

media file name for an annotation will bring you to the audio listen page (sec 3.1.3) for that media, with the available annotations shown. (This is supposed to move the audio player’s cursor directly to the moment of the note, but doesn’t always work – you may need to click the annotation’s timestamp after the listening page has loaded to move the audio player to the correct time.)

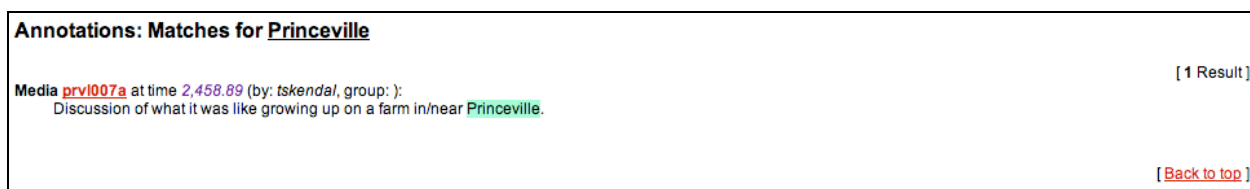


Figure 3.4.1.4 Annotation search results for “Princeville”

3.4.2. Variable extraction and coding (tabulation)

SLAAP has a number of tools build to enhance variable extraction and coding, what we typically call “variable tabulation” in the Wolframian variationist framework. Only user accounts with access to “tabulation” features have access to the pages described in this section. If you have access to the variable tabulation features, you will see a “Tabulation Summary” link at the top right of the main library screen (section 3.1.1), a “Tab” link at the top right of the listening (section 3.1.2) and downloading (section 3.1.3) screens, and you will have links to tabulations from each full record page (section 3.1.1). Figure 3.4.2.1, an extract from the full record for an interview in the DC Adolescents Project, illustrates the links to tabs in the full record view.

To create a new variable tab sheet, simply navigate to the variable tabulation page by way of any of the links described above. Then, select the variable you wish to work on. If a tab sheet does not exist for that variable, you will be prompted to create one. If a tab sheet does exist, you’ll be taken to that tab sheet.

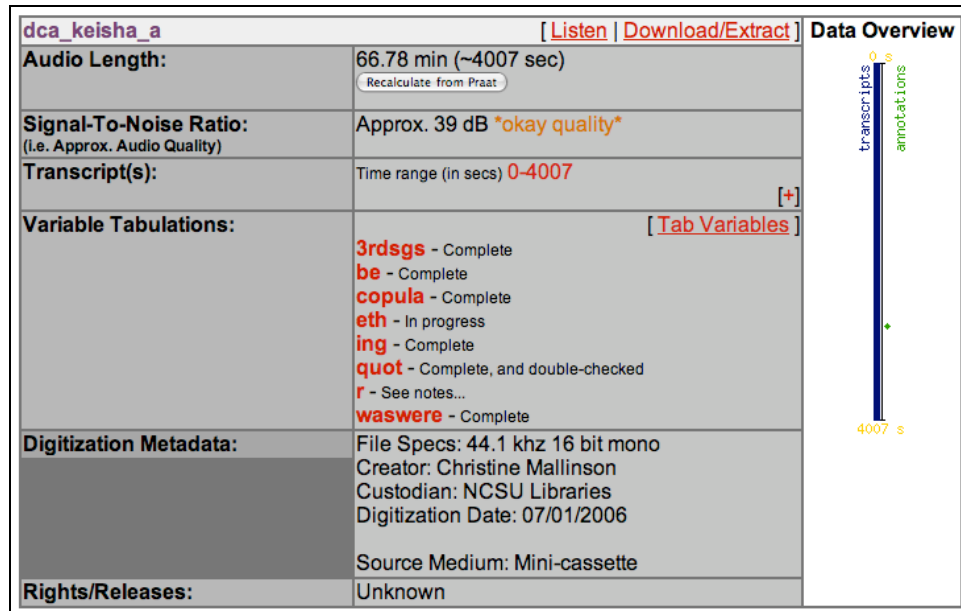


Figure 3.4.2.1 A media record from a DC Adolescents Project interview, showing tabs

****Importantly:** Currently, there can only be one tab sheet per variable per media file. That is, you cannot create a new tab sheet for a media file if that variable already exists for that file. Relatedly, anyone with tabulation-level access, and access to a given interview can access (and edit) the same variable tab sheet. Also, at present users cannot create new types of tab sheets (“variable definitions”). If you want/need to tab a variable not in SLAAP’s variable popup menu, Tyler can create the variable definition with you – just let him know. Typically, members of the NCLLP create variable definitions in SLAAP by brainstorming the needed and relevant features for that variable. The features you may be interested in for a given variable, say *copula deletion*, may not be the exact same as the features other SLAAP users are interested in for that variable. Yet, it benefits everyone to have an agreed upon set of features coded for, so we try to maximize our agreement along these lines. Even if it means you end up coding for a feature or two more than you’re interested in, it ensures that your data can be more easily compared to other relevant work.

Figure 3.4.2.2 shows a screenshot of the tabulation tool for the variable *copula absence* for one of the Princeville interviews. “Tabulation sheets” in SLAAP are stored in the database and accessed via a tool connected to the audio player. In addition to coding the relevant aspects of each variable realization, users also record the exact timestamp of the variable, allowing for

the easier review of variable tabs (this also enables many of the graphical transcript options that overlay variable tabs). In addition to recording the specific features of interest for each variable, through this tool you always record the speaker for a given variable realization, the variable's matrix context, and your level of confidence in your extraction and coding. The variables in Fig. 3.4.2.2 are all coded as “conf” (“confident”) but you can also specify “not confident” and “very confident”, making it easier to review your work later, and for your colleagues and/or research advisor to review and understand your work. The “DC” checkbox indicates “don't count” – you should use this to mark tokens that are noteworthy, but not appropriate “counted” cases for the variable at hand (for more about variable tabulating generally and “don't count” forms, see Blake 1997, Wolfram 1993). Note that you can tabulate variables for multiple speakers in the same tab sheet. Be sure to specify the correct speaker. You can also tab data on interviewers; speakers are ordered first in the “Speaker” popup menu, followed by the interviewers, who are enclosed in parentheses.

Audio Interface: prv007aa.wav (as mp3 for listening; length: 2,761.1 sec) [[Tabulation Summary](#) | [Full Record](#)] [[Download](#) | [Listen](#)] [[Tab](#)]

Interface:

Play (p) Stop (s) >> 1 Sec (f) << 1 Sec (b)

Get Time Move to Time Cursor: 756.78666 sec

Zoom: Zoom level (s): 1 2 5 10 Center to time: [] Set to Cursor (a) Zoom In

Tabulate: [[Create Tab-Delimited File](#) for downloading] [Variable: Copula Absence]

[Offset cursor by 2 seconds] [stored in tb_prv007aa_copula]

Tab#	Time	Speaker	Context	Form	Prec. Env.	Is/Are	Foll. Env.	Confidence	DC
3	576.70	PEO	the water 0 risin'	abs	NP	is	Ving	conf	
2	733.28	PEO	work for is positive	full	NP	is	NP	conf	dc
4	754.00	PEO	most people are homeless	full	NP	are	NP	conf	
5	755.00	PEO	when they 0 in	abs	pronoun	are	locative	conf	

* To Cursor: PEO Abs. NP Is NP Confident

Store Tab

Notes: first 10 minutes are mostly past tense...

Store Notes & Status Reset | Status: In progress

Figure 3.4.2.2 Example of the variable tabulation page

To edit an existing tab, click the tab number. You cannot delete tabulations. You can mark unwanted tabs as “DC” or you can reuse a tab number for an altogether different token if

you want to entirely remove a tab from the record. As shown in Fig. 3.4.2.2, tabs are ordered by their timestamp, not their tab number.

In addition to the play and stop buttons, there are buttons that move the audio player’s cursor back and forward one second. All of these buttons can be accessed via keyboard shortcuts using ctrl + the letter in parentheses – so “ctrl-b” for example moves the cursor back 1 second. There is also a “zoom” feature available that makes it easier to repeatedly listen to difficult tokens. This is described in section 3.1.3 about the listening page.

Notes about individual tabs or about the tab sheet or recording in general can be recorded in the “Notes” field. You should also use the “Status” popup menu to track the status of your variable tabulation sheets.

At any time you can download the tabulation data directly to your local computer, by clicking the “Create Tab-Delimited File” link. When you click this link, the tabulation sheet will reload and that link will be replaced with a new link “Download Tab-Delimited File”. You can download this data by right-clicking the link and selecting the appropriate option from your browser’s popup menu. This downloaded file will be a text-only file in a tab-delimited format. You can then open this file in Microsoft Excel, R, SPSS, or any number of programs.

[Review: Copula Absence for DC Adolescents Project]

14 tabulation sheets are currently available to you for Copula Absence for DC Adolescents Project (SLAAP has 23 total tab sheets for Copula Absence) [Examine cohort data]. [Change to Expanded View]

Variable	Interview/Media [sort]	Project	Status [sort]	Speaker(s) & Summary Analysis
Copula Absence	dca_alayn / dca_alayna_a	DC Adolescents Project	Complete	Alayna [breakdown]: abs = 29 (24%); ctr = 70 (59%); full = 20 (17%); N = 119; (dc = 42)
Copula Absence	dca_asia / dca_asia_a	DC Adolescents Project	Complete	Asia [breakdown]: abs = 21 (44%); ctr = 15 (31%); full = 12 (25%); N = 48; (dc = 13)
Copula Absence	dca_asia / dca_asia_b	DC Adolescents Project	Complete	Asia [breakdown]: abs = 4 (31%); ctr = 3 (23%); full = 6 (46%); N = 13; (dc = 1)
Copula Absence	dca_calan / dca_calandra_a	DC Adolescents Project	Complete	Calandra [breakdown]: abs = 22 (71%); ctr = 6 (19%); full = 3 (10%); N = 31; (dc = 5)
Copula Absence	dca_elisa / dca_elisa_a	DC Adolescents Project	Complete	Elisa [breakdown]: abs = 24 (73%); ctr = 4 (12%); full = 5 (15%); N = 33; (dc = 1)
Copula Absence	dca_grace / dca_grace_a	DC Adolescents Project	Complete	Grace [breakdown]: abs = 43 (53%); ctr = 31 (38%); full = 7 (9%); N = 81; (dc = 0)
Copula Absence	dca_grace / dca_grace_b	DC Adolescents Project	Complete	Grace [breakdown]: abs = 3 (30%); ctr = 6 (60%); full = 1 (10%); N = 10; (dc = 0)
Copula Absence	dca_keish / dca_keisha_a	DC Adolescents Project	Complete	Keisha [breakdown]: abs = 118 (86%); ctr = 9 (7%); full = 11 (8%); N = 138; (dc = 0)
Copula Absence	dca_latan / dca_latania_a	DC Adolescents Project	Complete	Latania [breakdown]: abs = 52 (68%); ctr = 22 (29%); full = 3 (4%); N = 77; (dc = 0)
Copula Absence	dca_shant / dca_shantell_a	DC Adolescents Project	Complete	Shantell [breakdown]: abs = 65 (63%); ctr = 30 (29%); full = 9 (9%); N = 104; (dc = 14)
Copula Absence	dca_shant / dca_shantell_b	DC Adolescents Project	Complete	Shantell [breakdown]: abs = 25 (64%); ctr = 11 (28%); full = 3 (8%); N = 39; (dc = 3)
Copula Absence	dca_shawn / dca_shawna_a	DC Adolescents Project	Complete	Shawna [breakdown]: abs = 19 (36%); ctr = 23 (43%); full = 11 (21%); N = 53; (dc = 0)
Copula Absence	dca_shiri / dca_shirlisa_a	DC Adolescents Project	Complete	Shirlisa [breakdown]: abs = 14 (17%); ctr = 47 (58%); full = 20 (25%); N = 81; (dc = 0)
Copula Absence	dca_shiri / dca_shirlisa_b	DC Adolescents Project	Complete	Shirlisa [breakdown]: abs = 13 (21%); ctr = 41 (67%); full = 7 (11%); N = 61; (dc = 0)

Figure 3.4.2.3 Summary overview of copula absence tabs for the DC Adolescents Project

The “Tabulation Summary” links (on the main library page and the tabulation pages) bring you to an overview list of all the tab sheets in the system available to your account. This is demonstrated in Figure 3.4.2.3, which shows *copula absence* tabs for the DC Adolescent Project. As illustrated in the figure, you can limit the overview to specific variables and specific projects. This summary overview page gives very simple absence and presence rates for most variables.

Clicking on a variable name (in the leftmost column) brings you to the tab sheet for that record. The “breakdown” link brings you to a more complete breakdown “analysis” for the tab sheet. This is shown in Figure 3.4.2.4.

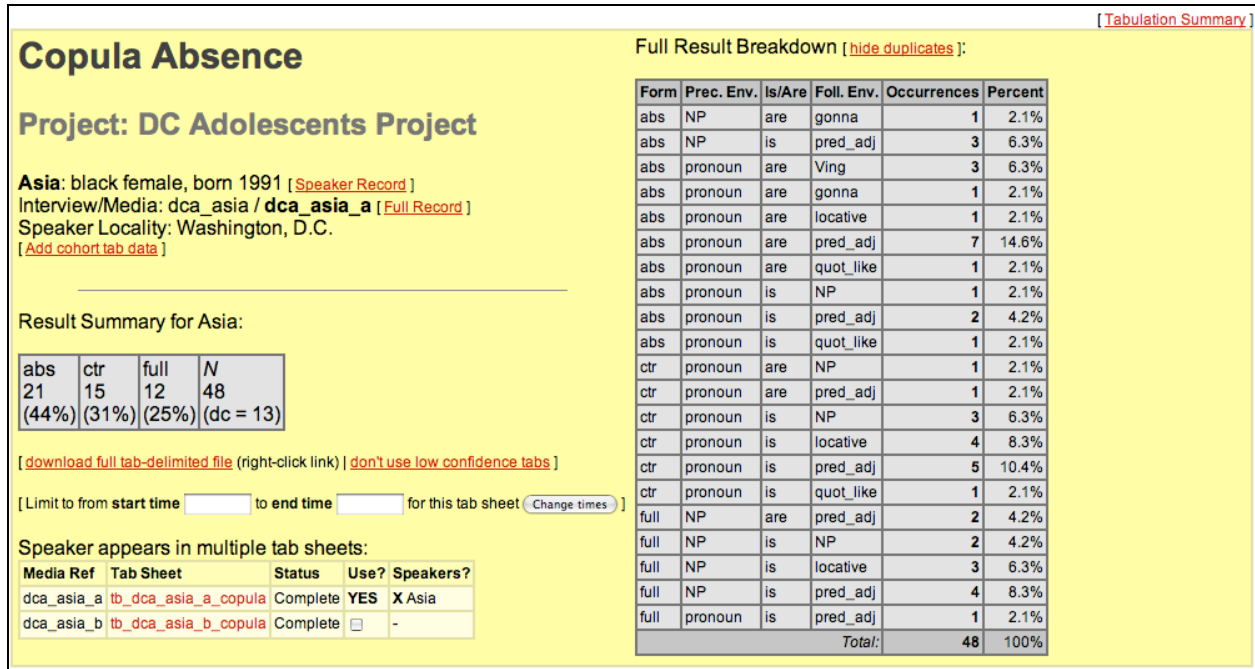


Figure 3.4.2.4 Tab breakdown for copula absence for a DC Adolescents Project media file

On this breakdown page, you can download the data in a tab-delimited format (as described above). You can also review the data with “low confidence” tabulations removed. Note that, since tab sheets are associated to media files (as opposed to full interviews or individual speakers), a given tab sheet may not contain all of the tabs for a speaker for the variable in question. In Fig. 3.4.2.4, for example, note that the bottom left of the screen declares that the “Speaker appears in multiple tab sheets” and that you can click checkboxes to add those additional tab sheets to the current breakdown.

There are not yet any features in place to conduct actual statistical analysis on variable tabulations within SLAAP. Features are planned that will better integrate with statistical analysis programs (such as R and GoldVarb). In the meantime, Tyler can help you quickly convert SLAAP tab sheet data into GoldVarb token files for variable rule analysis. The tab-delimited

files extractable from SLAAP readily work with Daniel Ezra Johnson’s (2008, 2009) Rbrul package for R.

3.4.3. Transcript-based speaker analysis tools

Analyses in Tyler’s dissertation (Kendall 2009) focus on variation in speech rate and pause and the relationships between speech rate and pause and “normal” variables. To conduct these analyses, he has built tools in SLAAP under the heading “Speaker Analysis”. If you have access to these features, you will see a “Speaker Analysis” link at the top of the main library screen. You will also have links to the speaker analysis features via the transcripts overview page (the identifiers in the “Speaker(s)” column will be links, see section 3.2.1).

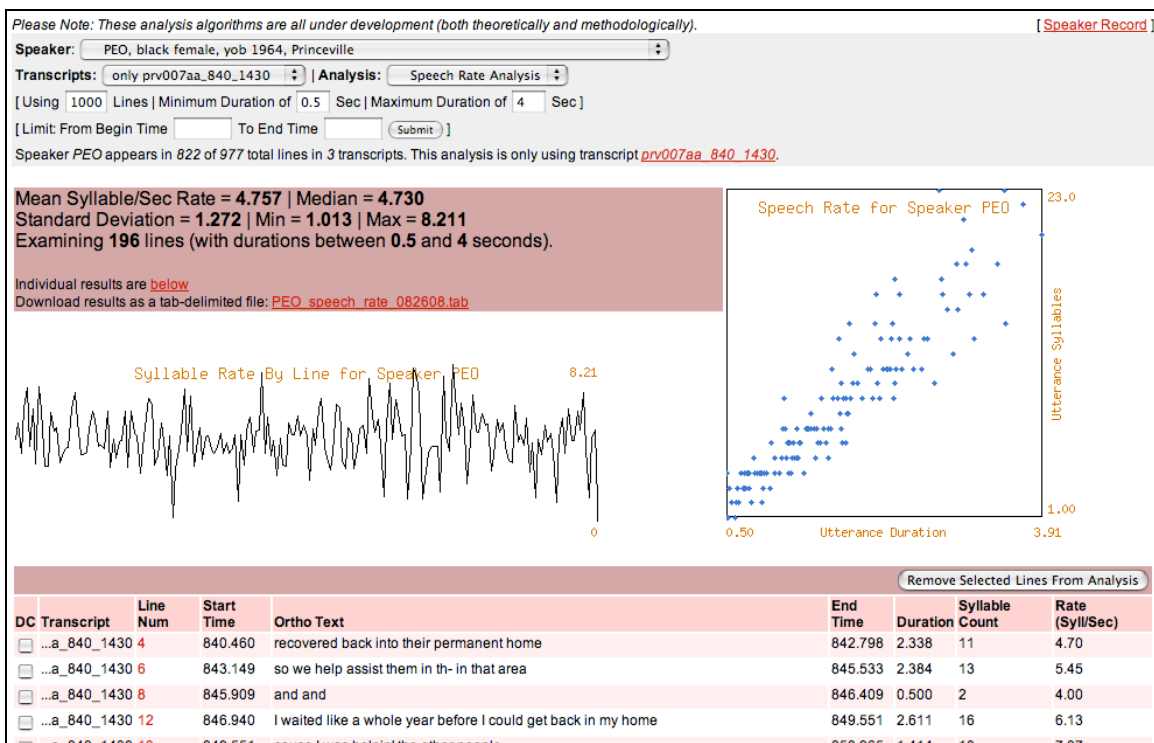


Figure 3.4.3.1 Speech rate analysis using one transcript for a Princeville speaker

Figure 3.4.3.1 illustrates the speech rate analysis feature and Figure 3.4.3.2 illustrates the pause analysis feature. There is also a pitch analysis feature, as well as a part-of-speech analysis

feature, on this page. Neither are shown here, and the part-of-speech analysis tool in particular has a long way to go before it's useful.

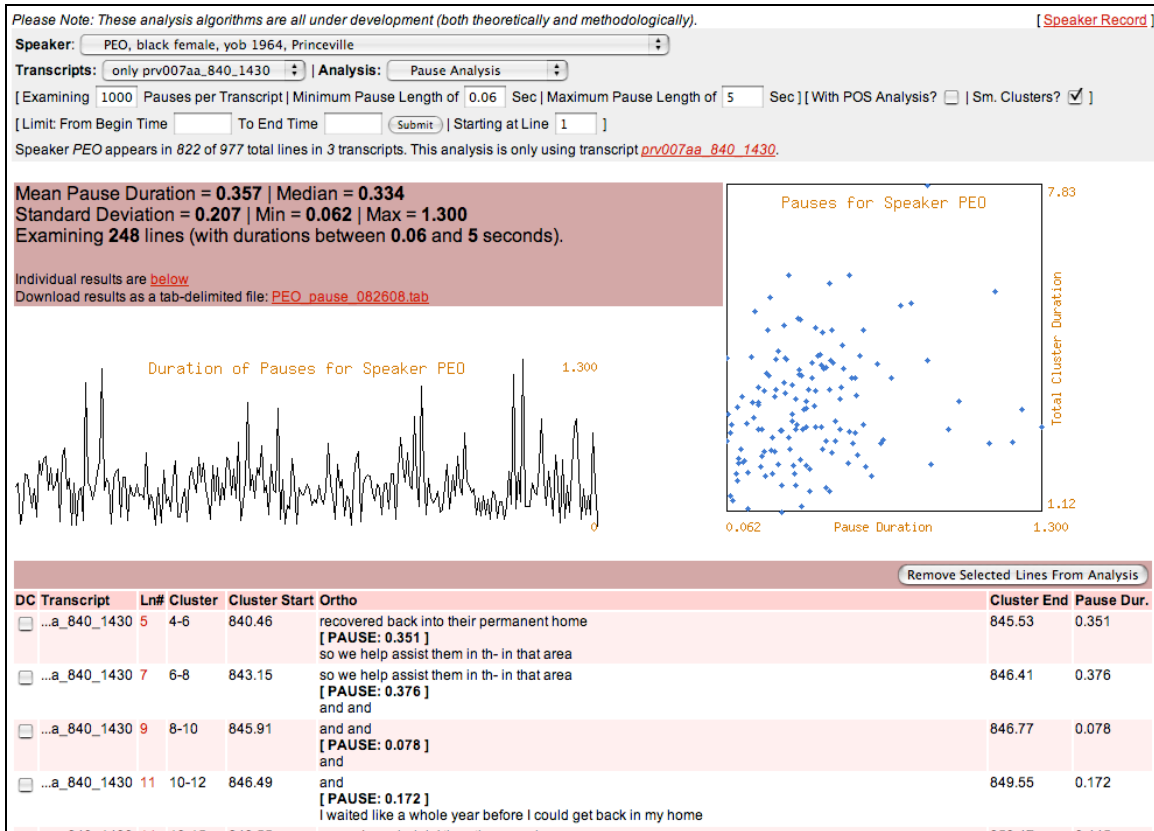


Figure 3.4.3.2 Pause analysis feature using one transcript for a Princeville speaker

3.4.4. Transcript-based lexical analysis tools

Users with access to the speaker analysis tools (section 3.4.3) also have access to the experimental transcript-based lexical analysis tools. If available to your account, these tools can be reached from the “Lexical Analysis” link at the top of the main library screen. Unlike the speaker analysis tools, which are oriented around each speaker, the lexical analysis tools are designed to aggregate across speakers within transcripts or projects. Most of the lexical analysis tools disregard temporal information and approximate more traditional corpus-oriented tools. As examples, Figure 3.4.4.1 shows a lexical frequency analysis for the transcripts in the Princeville

project; Figure 3.4.4.2 shows a section of the page that plots the lexical growth over token and time for each of the Princeville transcripts.

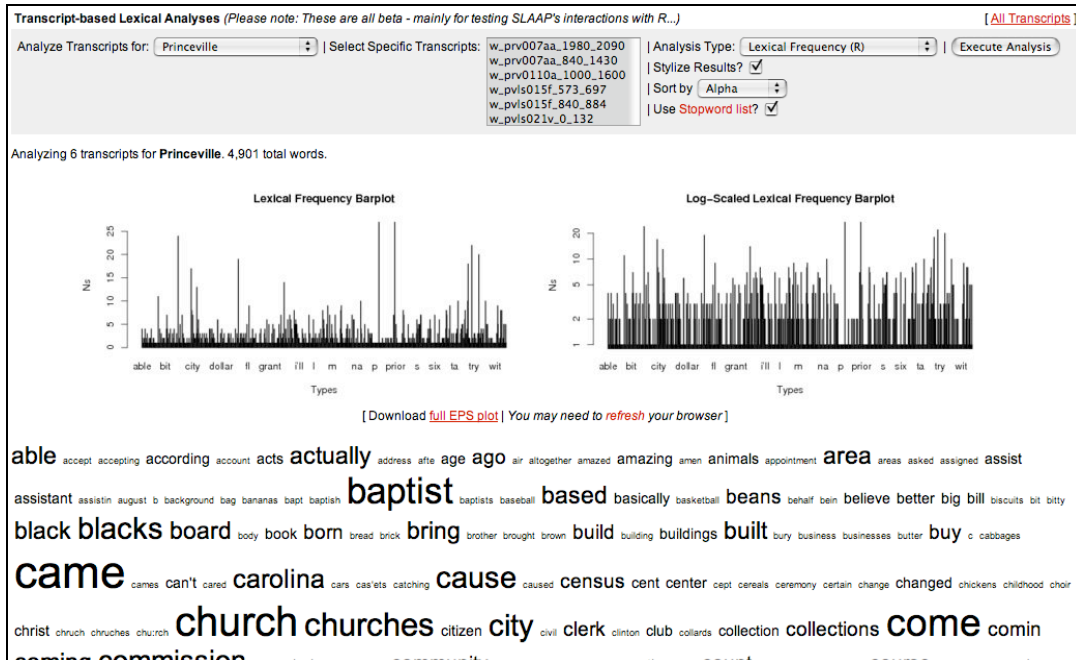


Figure 3.4.4.1 Lexical frequency analysis for Princeville transcripts

To a large degree, the lexical analysis features have been developed as ways for Tyler to practice building R (the statistical package and language, R Development Core Team 2007) into the SLAAP software. If you are interested in using or expanding these features, please talk to him.

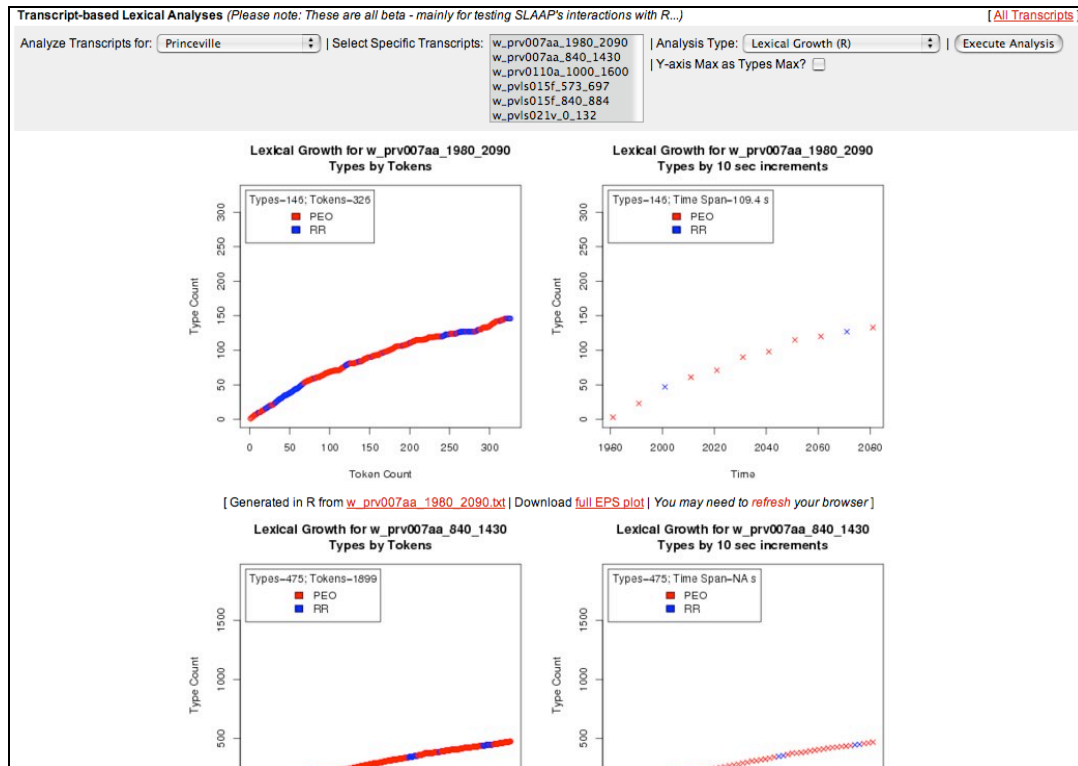


Figure 3.4.4.2 Lexical growth plots for Princeville transcripts

4. Digitizing audio cassettes and adding data to SLAAP

(see/attach White & Cullinan document)

5. Transcribing

This section provides guidelines for and outlines the transcription process, using Praat, for SLAAP. Section 5.1 focuses on Praat, and our recommended transcription conventions. Section 5.2 focuses on how to add Praat transcripts to the SLAAP archive.

Before proceeding with a transcript for a media file in SLAAP, be sure to confirm that that transcript does not already exist, or that segments of the media file that overlap with your segment have not already been transcribed. If there is already transcribed data in SLAAP that overlap, or abut, the segment you are interested in. Please export the current transcript (see

section 3.2.5) and then build on the existing transcript. Then, instead of following the instructions in section 5.2 (on uploading transcripts), contact Tyler about adding the replacing the current, smaller transcript, with your new version.

5.1. *Transcribing in Praat*

Although Praat (cf. <http://www.fon.hum.uva.nl/praat/>; Boersma and Weenink 2007) was not designed to be a transcription tool, it is quite possibly the best tool for creating time-aligned transcripts. Praat is free, open-source, and works on all the major computer operating systems¹. It also allows for multiple levels of annotation or transcript, beyond those used here (cf. Vaughn ms). Finally, and most importantly, Praat allows for arbitrarily fine-grained time-alignment for any number of speakers.² Other software specifically designed for transcribing (namely, Transcriber, <http://trans.sourceforge.net/>) share many of these features, but do not allow for the level of time-alignment capable within Praat³.

This section attempts to give a quick, but technically thorough overview of the transcription process in Praat, with transcription conventions for transcripts that intend to be added to the SLAAP archive.

5.1.1. Praat, the TextGrid, and TextTiers

Transcription in Praat takes place using the TextGrid object type. A transcript is made in a TextGrid object and each speaker in the transcribed interaction has a tier (a TextTier) of the TextGrid. Figure 5.1.1.1 shows the Praat editor window for transcript with two speakers, along with the audio data. Typically, the tiers are ordered by interest and amount of talk. So, the top tiers will contain the main interviewee(s); below those will be the interviewer(s), and at the bottom less talkative or important speakers as well as interlopers.

¹ Of course, other transcription software programs available (such as Transcriber, <http://sourceforge.net/projects/trans/>) are also free, open-source, and cross-platform. However, they do not provide the fine-grained time-alignment possibilities enabled by Praat.

² In fact, we argue that Praat is an excellent tool for transcription for everyone, beyond SLAAP users. A public tool is available on the SLAAP website – http://ncslaap.lib.ncsu.edu/tools/praat_to_text.php – that converts from Praat format to readable text to aid non-SLAAP users.

³ Transcriber's main problem here is that it does not allow partial speaker overlap. Utterances are organized by "turns" in Transcriber and two speakers' utterances can be adjacent or can completely overlap, but cannot partially overlap.

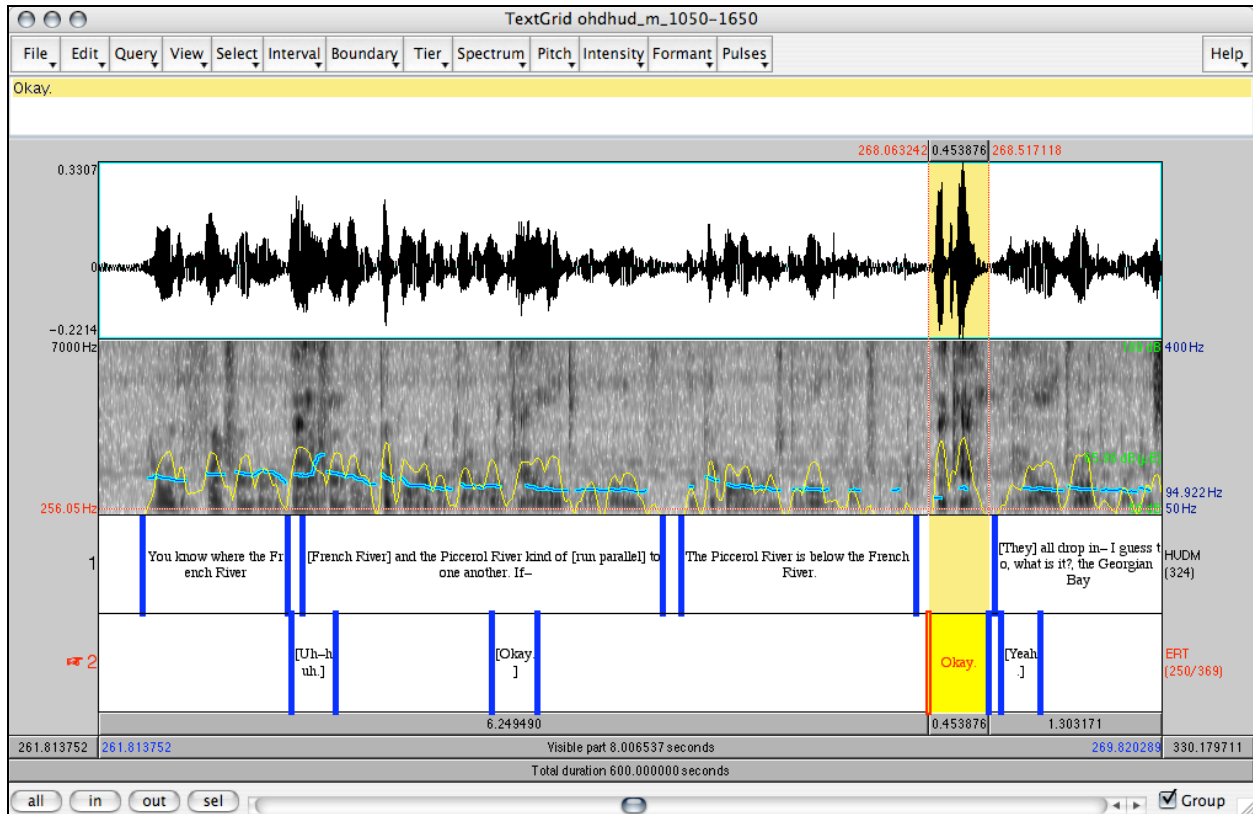


Figure 5.1.1.1 Screenshot of Praat Editor window (media: ohdhud_m)

To create a TextGrid for a new transcription, open the sound file for the passage you wish to transcribe, using the “*Read from file...*” menu command. The sound file will now appear in the object list and will be selected. Next, click the “*Annotate –*” button and select “*To TextGrid...*” (this is illustrated in Figure 5.1.1.2).

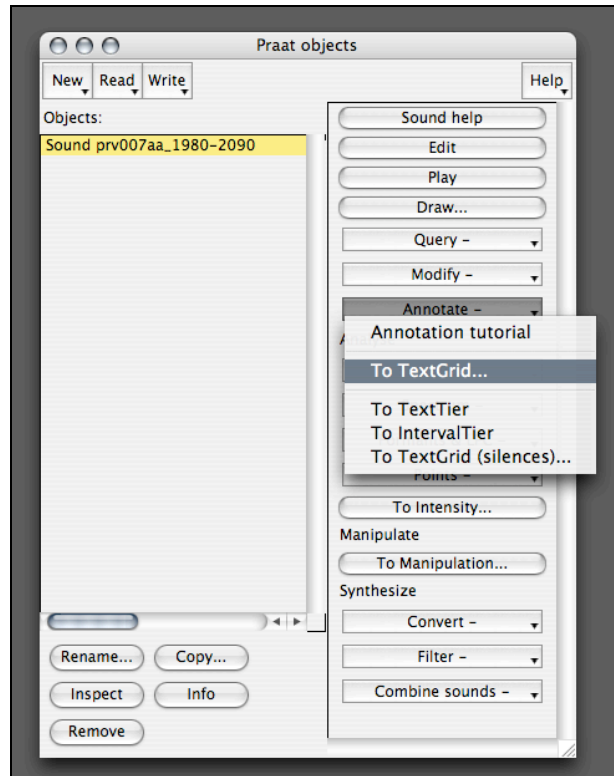


Figure 5.1.1.2 Screenshot from Praat, showing the creation of a TextGrid for a media file

A dialogue box will appear. In this dialogue, you need to replace the default text in the “*All tier names:*” field (typically “*Mary John bell*”) with the identifier for each speaker in your transcript (customarily each speaker’s initials or a (short) pseudonym) separated by spaces. You won’t be needing *point tiers* so should delete all the text in the second field.⁴ Sometimes you may not know all of the speakers (or how many speakers there are) until you are transcribing. That is okay; you will be able to add new tiers later if necessary⁵. For now, just enter the identifiers for the speakers you know about. Once you click “*OK*” a new TextGrid object will show up in your object list with the same name as the sound file. Select them both (by dragging over both, or by holding the “*shift*” key while clicking on the sound file). Then click the “*Edit*” button. You should now have a window that looks remotely like Figure 5.1.1.1 (above), but that does not yet have any transcription in it. You will also most likely need to zoom in, in order to

⁴ This is not intended as an introduction to the Praat software itself. Beginners are referred to the Praat website – <http://www.praat.org/> – for documentation and tutorials on using Praat. The Praat software’s help section and built in manual are also good resources. Beginning transcribers are urged to read Section 7 of the manual, on “*annotation*” – http://www.fon.hum.uva.nl/praat/manual/Intro_7__Annotation.html – for further guidance as to how to create and edit TextGrids. Selecting the “*Annotation tutorial*” from the “*Annotate -*” button-menu will also bring up the help documentation on annotation.

⁵ You can add a new tier in editor window by selecting the “*Add interval tier...*” from the “*Tier*” menu.

see a spectrogram. We'll address this in section 5.1.4. If you want to open an existing TextGrid transcript, instead of creating a new one, you can simply add that TextGrid to the object list by using the same “*Read from file...*” menu command that opened the sound file.

Importantly, in a Praat-based transcript, each tier of the TextGrid will provide a complete accounting for a speaker's speech over the course of an interview. Empty intervals indicate when speakers are silent, while text in intervals provides an orthographic representation for the speakers' talk during the time span of the interval.

So, how does one do this? And, what goes in those tiers? These are topics of the next section.

5.1.2. Delimiting the utterances

The most important aspect of a time-aligned transcript is the accurate delimitation of the speech timing. All pauses greater than about 60 milliseconds should be delimited as pauses, by being marked off with interval boundaries and **with no text** within the interval. In other words, boundaries should be placed at the edges of each phonetic utterance, and should delimit all silences by the speaker larger than about 60 ms.

Praat provides a number of ways to create interval boundaries. Once you identify the location of a boundary, by clicking on the waveform or spectrogram, the easiest way to create the boundary is to click the small circle that appears along the cursor-line at the top of each TextTier. You can also select the “*Add on [x]*” command from the “*Boundary*” menu (of course, these have keyboard shortcuts as well). To move boundaries, you can simply click and drag them. To erase an unwanted boundary, you can click it to select it (the selected boundary is always indicated in red), and then use the menu command “*Remove*” from the “*Boundary*” menu.

5.1.3. Orthographic conventions

To enter text into a TextGrid interval, you can simply type.

The primary utility of a time-aligned transcript – at least within its conceptualization within SLAAP – is to act as a proxy to the original recording, to allow for easy searching and browsing of the recording. It is not to make an exact, textually accurate representation of the speech (if that’s even possible). Along these lines, the orthographic transcript should use simple orthography and standard-like spelling. As a general rule, morphosyntactic variants (e.g., *was* for *were*) should be transcribed, but phonological variants (such as *r*-lessness, or Northern Cities vowel qualities) should not be⁶.

At the same time, the transcript text should accurately account for all the “noises” of speech, such as laughter, filled pauses (like “uh” or “um”), and restarts (e.g., “I- I- I di- didn’t mean to”) and misspoken words (e.g., “brack in the seventies”). Standard-like punctuation should be used, with the hyphen, -, used to indicated lexical and morphsyntactic restarts, as well as incomplete intonation. Silent pauses (of course!) should not be described or coded, as they are represented in all cases, by empty intervals.

Since a number of speech sounds (e.g., mutterings like “Mm-hm”) do not have codified spellings and some extremely common productions do have widely agreed upon non-standard written forms (e.g., “kinda” and “I’m’a”), Table 5.1.3.1 gives orthographic guidelines and examples for some commonly encountered “words”.

Table 5.1.3.1 SLAAP spelling conventions, examples

Uh-huh	Uh-uh	Gonna
Uh-hum	Okay	I’m’a
Mm-mm	Mkay	Wanna
Mm-hm	Nyah	Kinda

It is also important to code for three features, which have special characters. These are described in Table 5.1.3.2. Additionally, transcribers are urged to include notes about their transcripts, and about interesting speech features found in the speech. The convention for this is also included in Table 5.1.3.2.

⁶ Of course, transcribers and analysts may be interested in describing pronunciation features – these can be incorporated into the transcript through the use of intra-linear notes (see (4) in Table 2). These notes are then parsed by the NC SLAAP system when the transcript is uploaded to the archive (see Section 3.5) and made available to users as inter-linear notes within the transcript.

Table 5.1.3.2 SLAAP transcription conventions, special characters

Feature	Special Chars	Example
<p>1. <u>Overlap</u>: Speaker overlap should be noted by the use of square brackets, for all parties to the overlap. The overlap markers should, however, only be placed at word boundaries. Since utterances are accurately time-aligned through the tier boundaries, these markers are aids for readers and not critical for timing determinations. Therefore, they do not need to be (highly) accurately placed.</p>	<p>[]</p> <p>e.g., A: So [I wen-] B: [You did] what?</p>	<p>RR: Habitat [for /Humanity/ PEO: [Habitat for] Humanity came</p> <p><i>prv007aa_840_1430, 94-95</i></p>
<p>2. <u>Unintelligible/inaudible speech</u>: Slashes should be used to enclose sections of unsure transcription. Transcribers can place “best guesses” within the slashes, or can write <i>/unintelligible/</i> for unintelligible talk or <i>/inaudible/</i> for inaudible talk. For unintelligible talk of less than three syllables, transcribers can also use question marks, ?, within the slashes to indicate each syllable of unintelligible speech.</p>	<p>//</p> <p>e.g., <i>/unintelligible/</i> <i>/inaudible/</i> <i>/??/</i> (= 2 syllables)</p> <p><i>/and the car/</i> <i>/and ? car/</i></p> <p><i>/PHONE RINGS/</i></p>	<p>“Here? /unintelligible/ High school.”</p> <p>(<i>bee0010a_0_2756, 1588</i>)</p> <p>“Well a rockweiler /really/ got”</p> <p><i>bee0010a_0_2756, 864</i></p>
<p>a. <u>Obscuration</u>: Slashes can also be used to obscure real names, when the transcription of someone’s real name is inappropriate or not allowed, or to replace real names with pseudonyms.</p>	<p>//</p> <p>e.g., <i>/NAME/</i> <i>/Alayna/</i> <i>/Center City/</i></p>	<p>“How old are you, /Alayna/?”</p> <p><i>dca_analyna_a_0_4245, 12</i></p>
<p>3. <u>Non-linguistic/meta-linguistic noises</u>: Noises like laughter, hand clapping, and throat clears should be indicated by short descriptions enclosed within angle brackets. These should only be used to describe actual noises, not features like voice quality or [laughter during speech].</p>	<p>< ></p> <p>e.g., <laugh></p>	<p>“I was a full-time mommy. <cough>”</p> <p><i>ptx1110a_300_904, 315</i></p> <p>ERT: <hah hah [hah]> CLHF: <[heh heh] heh heh heh></p> <p><i>ohdclh_f_991_1453, 275-276</i></p>

<p>4. <u>Line-level notes:</u> Notes can be included by the use of parentheses in a transcribed utterance. Features like voice quality or interesting pronunciations are appropriate to note this way. Parenthesized notes in a Praat TextGrid will be parsed by the NC SLAAP software and converted to inter-linear notes upon uploading.</p>	<p>() e.g., (whispered) (while chewing) (tape clicks) (window=winder)</p>	<p>“football games, that sort of /stuff/. (loud talking in the background!)” <i>dca_calandra_a_0_2500, 1712</i> “Because that's where my grandfather is and my mom she live- (creaky on 'live')” <i>dca_calandra_a_0_2500, 2353</i></p>
--	--	---

5.1.4. Recommended Praat settings

Transcribing in Praat is a time-consuming activity – all transcription is. However, with some practice and some customization of Praat’s settings, the transcribing process can be streamlined (as well as made more accurate). Table 5.1.4.1 lists a few of the Praat settings the authors find most useful (Figure 5.1.1.1 illustrated Praat with these settings).

Table 5.1.4.1. Recommended Praat settings for transcription

<p><u>Zoom level/window size:</u> An 8 second window in the Praat Editor seems to be the best resolution to readily see the short pauses that should be delimited with interval boundaries.⁷ Smaller windows are not recommended, as speech features like voiceless obstruents will look like interruptions in the phonetic utterance.</p>
<p><u>Spectrogram/analysis settings:</u> Try setting the spectrogram display to show from 0 Hz to 7000 Hz for transcribing⁸. The higher frequencies, while typically not viewed for sociophonetic work, can help you see utterance-final frication, aspiration, and stop release, which should be included in the utterance’s interval. These features can be hard to hear in noisy signals. We also recommend enabling (showing) pitch and intensity analyses in the spectrogram window⁹. These help in the identification of pauses. Pitch can also help you see which speaker an utterance belongs to, especially when the speakers have different pitch ranges.</p>

As a further recommendation, it is advised to transcribe in Praat on the fastest available computer. Especially, for longer sound files, Praat can be slow to load a window of sound on

⁷ This can be “eye-balled” by selecting 8 seconds of speech and “zooming to selection” from the “View” menu or “sel” button, or it can be done by selecting the “Zoom...” option from the “View” menu and entering an 8 second time range.

⁸ This is set from the “Spectrogram settings...” option in the “Spectrum” menu.

⁹ These may be “on” by default. Pitch is typically displayed as a blue line in the spectrogram window; intensity is displayed as a yellow line. You can enable/disable these via the “Pitch” menu and the “Intensity” menu respectively.

older computers. It can be extremely frustrating and time-wasting to have to wait 5 to 10 seconds each time you scroll to the next 8 seconds of audio to transcribe.

5.2. Adding Praat transcripts to SLAAP

Once a Praat transcript is complete (for the time range of interest), a SLAAP user with the proper permissions can add the transcript to the archive in a few relatively straightforward steps. If you don't have access to the upload page, contact one of the authors, who will upload it for you or give you access to this feature.

5.2.1. Steps for adding Praat transcripts to SLAAP

1. In Praat, from the *Praat Object* window, save the TextGrid file as a **chronological** text file by selecting that option from the “Write...” menu. SLAAP only parses chronological text files, and will give the user an error if the TextGrid is in the “normal” format.

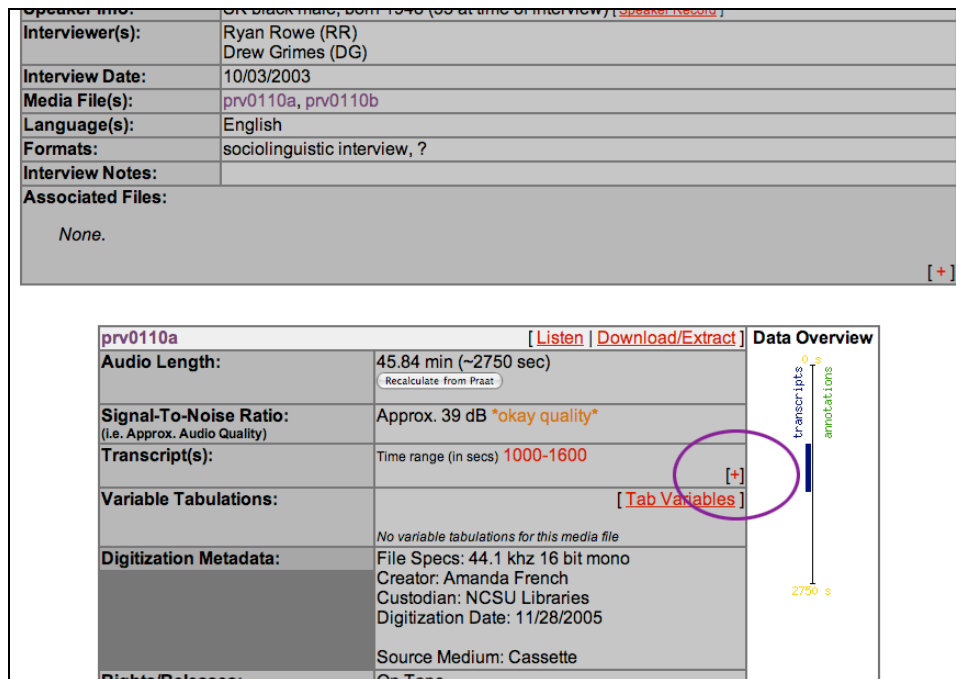


Figure 5.2.1.1 Partial screenshot of the full record view, highlighting the add transcript link

2. In SLAAP, navigate to the full record view (see section 3.1.2) for the transcription’s interview. Locate the correct media file and click on the “+” link in the section labeled “Transcript(s)”. This is circled in Figure 5.2.1.1, a partial screenshot of the full record view for a Princeville interview.
3. The next screen, shown in Figure 5.2.1.2, is admittedly overly verbose and needs to be improved. For the time being, the only two parts that you really need pay close attention to are the field labeled “Start Time” and the “Choose File” button. Both of these are highlighted in the figure. You should also check the appropriate box if this transcript covers the complete media file. Be sure to read the fine-print following the “Start Time” field – if you are transcribing an excerpt of a sound file and you extracted the excerpt outside of Praat, you likely need to enter the time (in seconds) that that excerpt starts at in relation to the overall sound file.

Transcript Upload

Please upload the transcript file (as a Praat TextGrid 'chronological text file' only)

Associated with **Media**: This is required to store the transcript in the database.

Associated with a **Project**: (this is to help pick speakers in the next screen)

File Type:

- Praat TextGrid (in 'chronological text file' format only)
- As a simple Text file (this will be stored with made up timestamps and marked as a temp transcript in the system)

Is the **complete** media file transcribed in this file? (check if yes)

If not complete, enter the **Start Time**: **seconds** (**IMPORTANT:** Your TextGrid is probably time stamped in relative time - that is, it probably starts at time 0 regardless of where in the audio file you transcribed. If so, manually enter the start time for this transcript. The parser will update the other times. Contact Tyler if you have questions about this.)

If not complete, and you are uploading a **Text File**, please enter the **End Time**: **seconds** (The end time of the media file will be used if you don't enter anything.)

Action to take?

- Parse (and Display)
- Store File Directly

Send this file:

Figure 5.2.1.2 Transcript upload screen, highlighting most important elements

4. After clicking the “Upload” button, you will be presented with a screen, such as that shown in Figure 5.2.1.3, that lets you review your uploaded transcript. Scroll through the entire transcript to confirm that the system parsed it correctly. If there are errors,

you'll want to abort the upload and revisit your transcript in Praat (or with a text editor). At the bottom of the page, are popup menus that you use to link the speakers in the transcript (regardless of how you labeled them in the transcript) to the actual speakers in the SLAAP system. If your transcript's identifiers match speakers' identifiers as registered in SLAAP, the system will attempt to select for you the correct speakers. Be sure to confirm the correct speakers are selected. On occasion you may have speakers in your transcripts that are not otherwise registered in SLAAP as speakers, such as interlopers who briefly appear in interviews and aren't recorded as "real" speakers in the interview. You can leave these speakers as the "[If an interviewee, select here]" option, which will be equivalent to "anonymous" for SLAAP's purposes. You can (and should) also record any summary information you like as well as notes. The authors recommend you record in the notes field information about the transcription process (such as whether your transcript is based on a previous Word document transcript).

Transcript Upload

File was uploaded successfully and stored.
 Reading transcriptfiles/prv0110a_1000_1600chron.TextGrid into an array for parsing.
 First line = "Praat chronological TextGrid text file"
 Transcript runs from 1000 to 1600 (2000 to 2600 after adjusting) and has 2 speakers.
 Speaker 0 is SK
 Speaker 1 is RR
 Built a Transcript with 877 lines.

Scroll to the bottom for storage options (if available).
 Transcript named: w_prv0110a_2000_2600.

This is what the transcript looks like:

Line	Start	Spkr	Text	End
1	[2000.00]	SK:		[2000.49]
2	[2000.00]	RR:		[2010.11]
3	[2000.49]	SK:	down this way um	[2002.32]
4	[2002.32]	SK:		[2003.49]
5	[2003.49]	SK:	to a /Mid Cross Row/	[2004.79]
6	[2004.79]	SK:		[2004.92]
7	[2004.92]	SK:	you could go to school	[2005.98]
8	[2005.98]	SK:		[2007.09]
...				
872	[2598.02]	RR:		[2598.65]
873	[2598.65]	RR:	the store [/that ?/]	[2599.74]
874	[2599.07]	SK:	[right]	[2599.42]
875	[2599.42]	SK:		[2599.80]
876	[2599.74]	RR:		[2600.00]
877	[2599.80]	SK:	/right.../	[2600.00]

Associate **speakers**:

SK:

RR:

*(If, for some reason, the correct speakers are not available, please **Add the speakers** and then start the transcript upload process over.)*

Summary (optional):

Notes (optional):

[Return to Upload page](#) if anything looks wrong (especially the time stamps)

Figure 5.2.1.3 The very top and bottom from a screenshot of the final transcript upload page

Note Tyler is aware of two common problems for the uploading of transcripts. First, transcripts with Unicode characters seem to fail to work with the transcript upload tool. Newer versions of Praat appear to support Unicode better than earlier versions, so this seems to be increasingly a problem. If the above steps fail for you, and you are sure you have saved your transcript in **chronological** format, please contact Tyler. Second, it is easy to accidentally include carriage returns (presses of the “return” key) in your TextGrid text, and not to notice

those returns. This is especially problematic if you are copy-pasting text into a Praat TextGrid from, say, a legacy Word document transcript. Those carriage returns will cause the parser to misread your document. You will see this when you confirm your transcript. It is demonstrated in Figure 5.2.1.4, where a hidden carriage return at the end of line 9 causes the following lines to be mis-parsed. If this happens, you should abort the upload and revisit your transcript with Praat.¹⁰

Transcript Upload

File was uploaded successfully and stored.
 Reading transcriptfiles/prv0110a_1000_1600chron-err.TextGrid into an array for parsing.
 First line = "Praat chronological TextGrid text file"
 Transcript runs from 1000 to 1600 (2000 to 2600 after adjusting) and has 2 speakers.
 Speaker 0 is SK
 Speaker 1 is RR
 Built a Transcript with 878 lines.

Scroll to the bottom for storage options (if available).
 Transcript named: w_prv0110a_2000_2600.

This is what the transcript looks like:

Line	Start	Spkr	Text	End
1	[2000.00]	SK:		[2000.49]
2	[2000.00]	RR:		[2010.11]
3	[2000.49]	SK:	down this way um	[2002.32]
4	[2002.32]	SK:		[2003.49]
5	[2003.49]	SK:	to a /Mid Cross Row/	[2004.79]
6	[2004.79]	SK:		[2004.92]
7	[2004.92]	SK:	you could go to school	[2005.98]
8	[2005.98]	SK:		[2007.09]
9	[2007.09]	SK:	up here but if you live on the other side of /Mid Cross Road/ to had to go to /Caneda/	[2009.90]
10	[1000.00]	:	1009.903007 1018.622206	[1000.00]
11	[1000.00]	:	1010.110652 1010.525944	[1000.00]
12	[1000.00]	:	1010.525944 1011.076204	[1000.00]
13	[1000.00]	:	1011.076204 1012.425901	[1000.00]
14	[1000.00]	:	1012.425901 1013.401836	[1000.00]
15	[1000.00]	:	1013.401836 1013.765215	[1000.00]

Figure 5.2.1.4 Example of a transcript upload error caused by extraneous carriage returns

6. Browser and software requirements

As is mentioned periodically in this text, there are some web browser requirements necessary for all of SLAAP's features to work correctly, including the QuickTime player and JavaScript. See <http://ncslaap.lib.ncsu.edu/requirements.php> for the full list.

¹⁰ You can also edit TextGrid files with a simple text editor. In fact, it's often easier to search for, and fix, these accidental carriage returns in a text editor than in Praat.

7. References and further reading

- Blake, Renée. (1997). Defining the envelope of linguistic variation: The case of “don't count” forms in the copula analysis of AAVE. *Language Variation and Change* 9. 57-79.
- Boersma, Paul and David Weenink. (2007). Praat: Doing phonetics by computer.
[<http://www.praat.org/>]
- Johnson, Daniel Ezra. (2008). Rbrul. [http://www.ling.upenn.edu/~johnson4/Rbrul_manual.html]
- . (2009). Getting off the Goldvarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3(1): 359-83.
- Kendall, Tyler. (2006-2007). Advancing the utility of the transcript: A computer-enhanced methodology. *Linguistica Atlantica* 27-28: 51-55.
- . (2007a). Enhancing sociolinguistic data collections: The North Carolina Sociolinguistic Archive and Analysis Project. *Penn Working Papers in Linguistics* 13.2:15-26.
- . (2007b). On the status of pause in sociolinguistics. Paper presented at the Linguistic Society of America 2007 Annual Meeting: Anaheim, CA. January.
- . (2007c). Automatic transcript summarization and keyword generation for the North Carolina Sociolinguistic Archive and Analysis Project. NC SLAAP Working Papers, March 5th, 2007. [<http://ncslaap.lib.ncsu.edu/pdfs/Kendall-autosummarizing030507.pdf>]
- . (2008a). On the history and future of sociolinguistic data. *Language and Linguistics Compass* 2.2:332-351.
- . (2008b). Identity, performance, and “consciousness”: The use of locally salient linguistic forms in a formerly isolated community. Paper to be presented at New Ways of Analyzing Variation (NWAY) 37: Houston, TX. November.
- . (2009). *Speech Rate, Pause, and Linguistic Variation: An Examination Through the Sociolinguistic Archive and Analysis Project*. Doctoral Dissertation. Durham, NC: Duke University.
- , and Amanda French. (2006). Digital Audio Archives, Computer-Enhanced Transcripts, and New Methods in Sociolinguistic Analysis, Paper presented at Digital Humanities (ALLC/ACH) 2006: Paris, France. July.

Ochs, Elinor. (1979). Transcription as theory. In *Developmental Pragmatics*, eds. Elinor Ochs and Bambi Schieffelin, 43-72. New York: Academic Press.

R Development Core Team. (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. [<http://www.R-project.org/>]

Wolfram, Walt. (1993). Identifying and interpreting variables. *American Dialect Research*, ed. Dennis Preston, 193-221. Amsterdam: John Benjamins.

8. Acknowledgements

The Sociolinguistic Archive and Analysis Project has been made possible with support from a number of organizations. We thank the North Carolina State University Libraries, the William C. Friday Endowment at NC State University, and the Duke University Graduate School for contributed funding. Much of the data stored in SLAAP were obtained through support by the National Science Foundation (including most recently grant number BCS-0542139).