The Value of Relational Databases for Time-Aligned Annotation

Tyler Kendall
Northwestern University &
North Carolina State University

t-kendall@northwestern.edu

Slides: http://ncslaap.lib.ncsu.edu/pdfs/aacl2009-reldb.pdf

AACL 2009, University of Alberta, Oct. 10th 2009

Time-Aligned Transcription

- Recent work in a number of linguistic disciplines has stressed the importance and utility of time-aligned annotation for linguistic corpora (cf. Bird and Liberman 2001)

- (It's not very controversial to claim that time-aligned annotation methods are important.)

XML v. (Relational) Databases

- At the same time, XML has emerged as a popular technology for structuring various sorts of corpus data (used by Text Encoding Initiative (TEI), Corpus Encoding Standard (CES), etc.), more so than any other data management technology (cf. Gries 2009, McEnery et al. 2006; e.g., Simons et al. 2004).

- With few exceptions (namely, Davies 2005), relational database engines, such as MySQL or PostgreSQL, have not been discussed in the linguistic literature as useful for the storage and manipulation of linguistic corpora.

Davies 2005

- "The advantage of using relational databases for large corpora" (*IJCL* 10.3)
  - Davies' paper focuses on the size-speed benefits of relational databases in particular with large (e.g., > 100 million word) corpora.
  - He outlines an approach which stores pre-compiled queries across the corpus data, such as n-grams and frequency tables
  - 4 goals:
    1. Size
    2. Levels of annotation
    3. Speed
    4. Modularity

Davies 2005, 2

- I agree, but don't mean for this talk to simply re-iterate that paper.
- I will focus on some other areas of advantage from a different perspective
  - Using smaller, spoken language datasets
  - Considering transcript data (and their linkage to their source audio) as the data of interest
    - Still though, the approach I am discussing centers on the storage of "the fundamental units of the corpus … in sequential rows of a database" (Davies 2005: 309).

Goals

- My goal today is two-part:
  1. To demonstrate and describe an approach to linguistic data management and time-aligned annotation based heavily on a database engine design
     - The **Sociolinguistic Archive and Analysis Project** (SLAAP; http://ncslaap.lib.ncsu.edu/)
  2. And to wonder aloud about a possible extension to broader areas of corpus design and sharing
     - I am interested in a broader conversation about the strengths and weaknesses of relational DBs v. XML
     - I hope to remember to disclaim later that I'm not particularly married to this approach – I mostly believe it raises some interesting ideas and I'd love to hear feedback…

The Sociolinguistic Archive and Analysis Project (SLAAP)

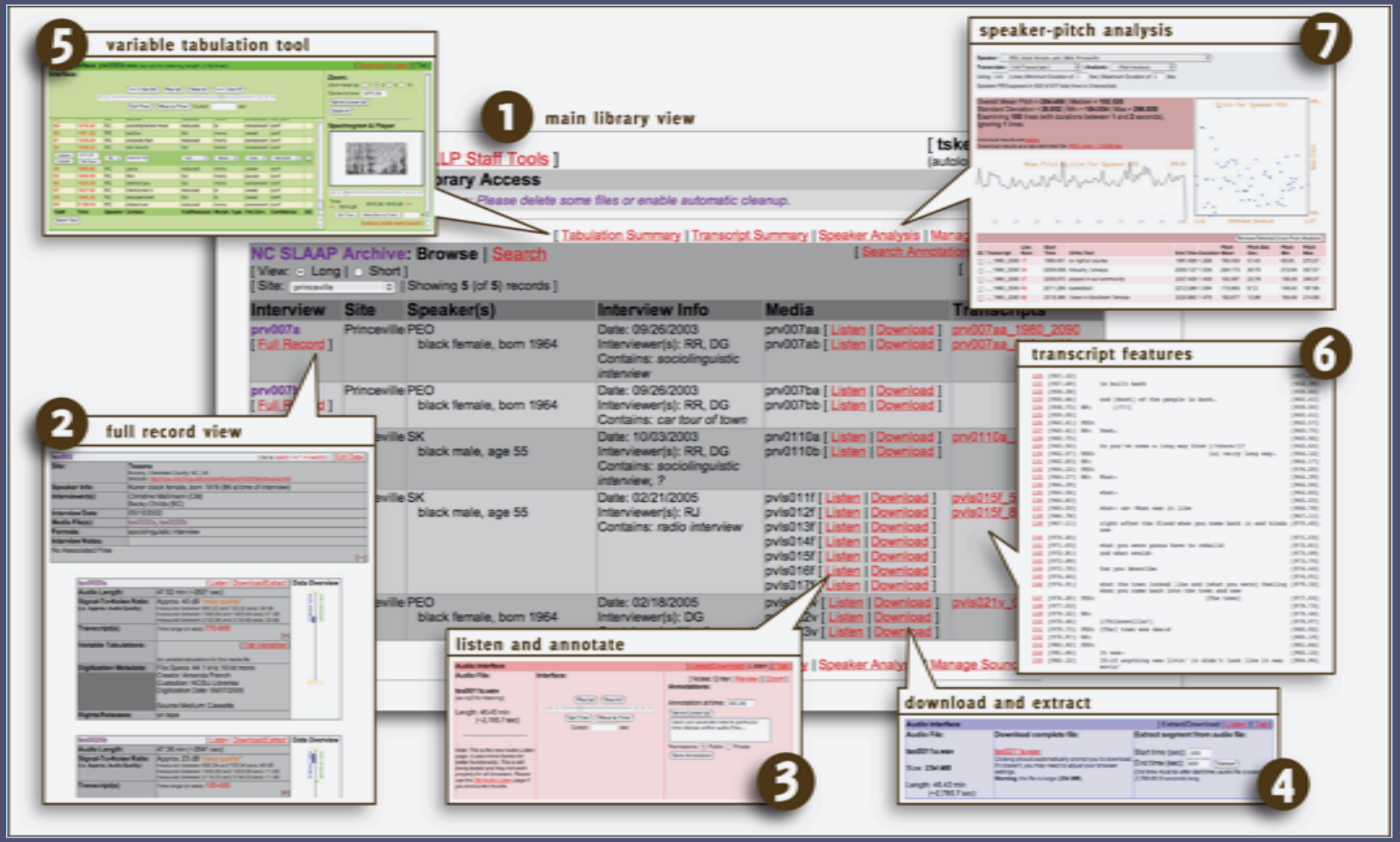
- An initiative at North Carolina State University to digitize a large sociolinguistic interview collection (and increasingly other collections) for preservation and accessibility
  - We are making the collection web accessible, so (with adequate permissions) scholars can access their data from anywhere in the world

- But SLAAP is more than an archive:
  - It is web-based software that seeks to enhance linguistic data through the development of analytic tools and data-models
  - Through this, we are exploring new, computer-enhanced techniques for interacting with the collection and for conducting sociolinguistic and linguistic analyses
  - In Poplack's (2007) terms, SLAAP is a **tool** with no projected **end-product**

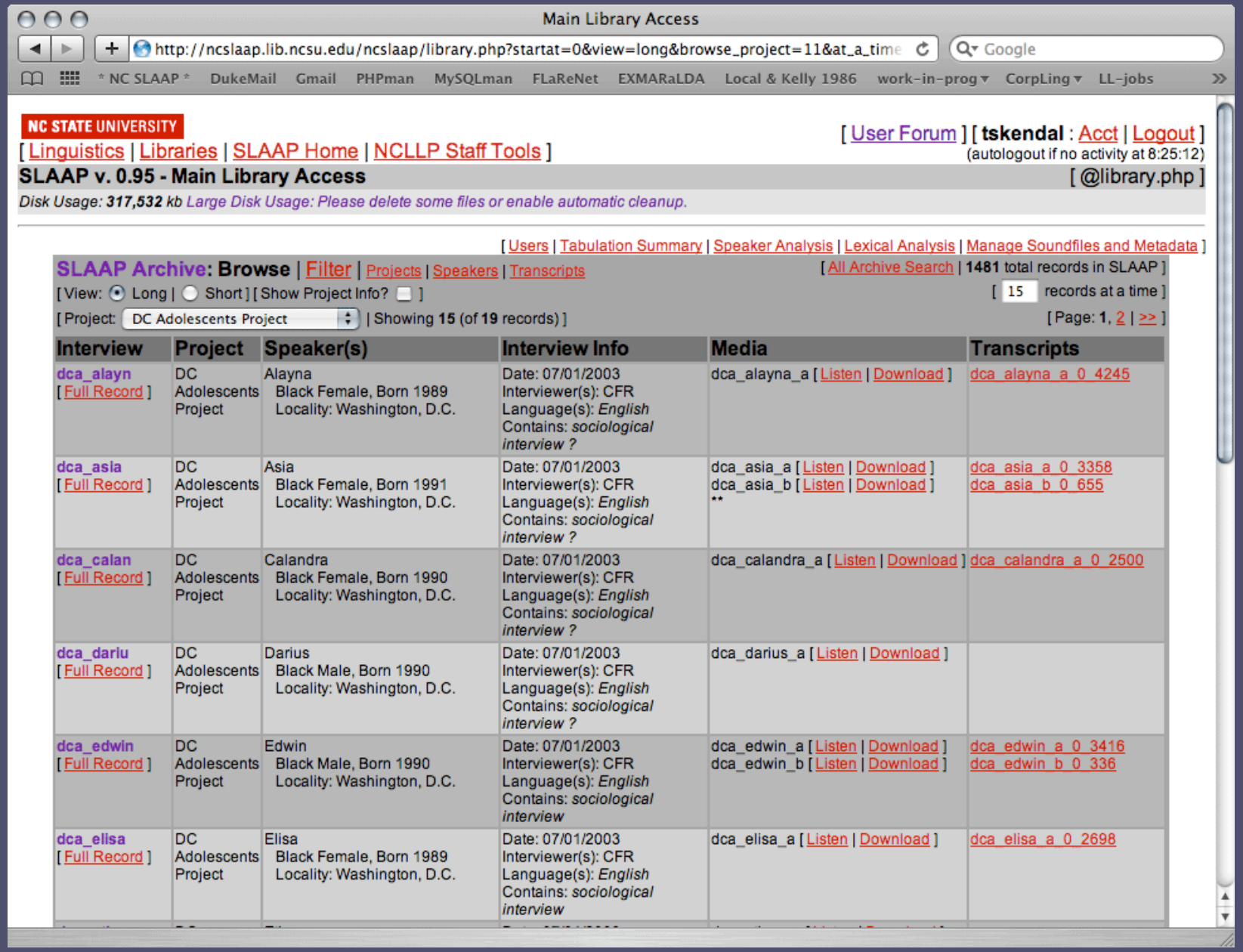
The (Current) Archive

- Currently (October 2009), SLAAP houses:
  - Over 1,450 interviews
  - Over 2,450 media files (> 1,125 hours of audio)
  - Not just from North Carolina…
- Transcription is ongoing (and slow):
  - Over 34 hours of time-aligned transcripts (~340,000 words)

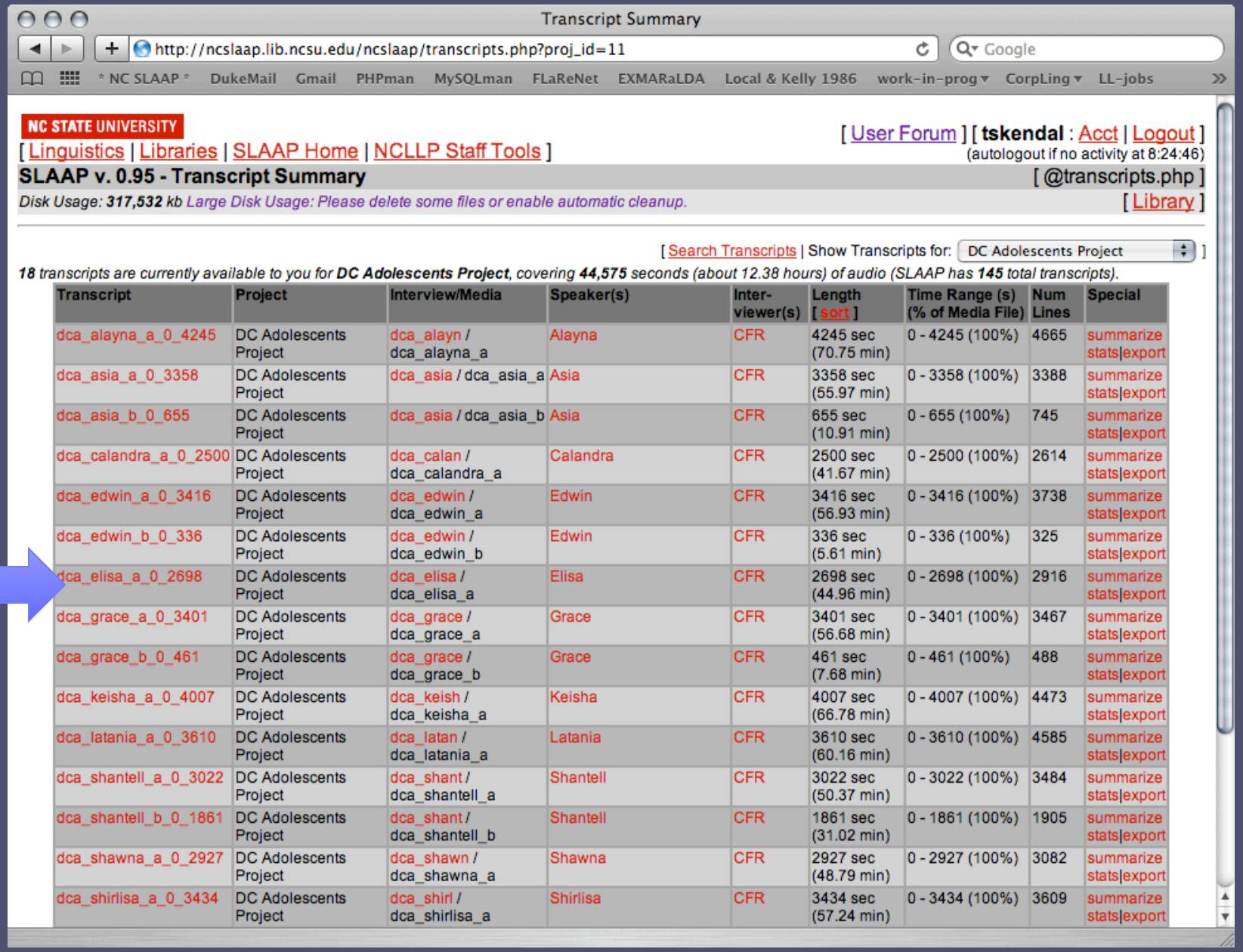
SLAAP: http://ncslaap.lib.ncsu.edu/



[Assorted screenshots from SLAAP, from Kendall (2007)]

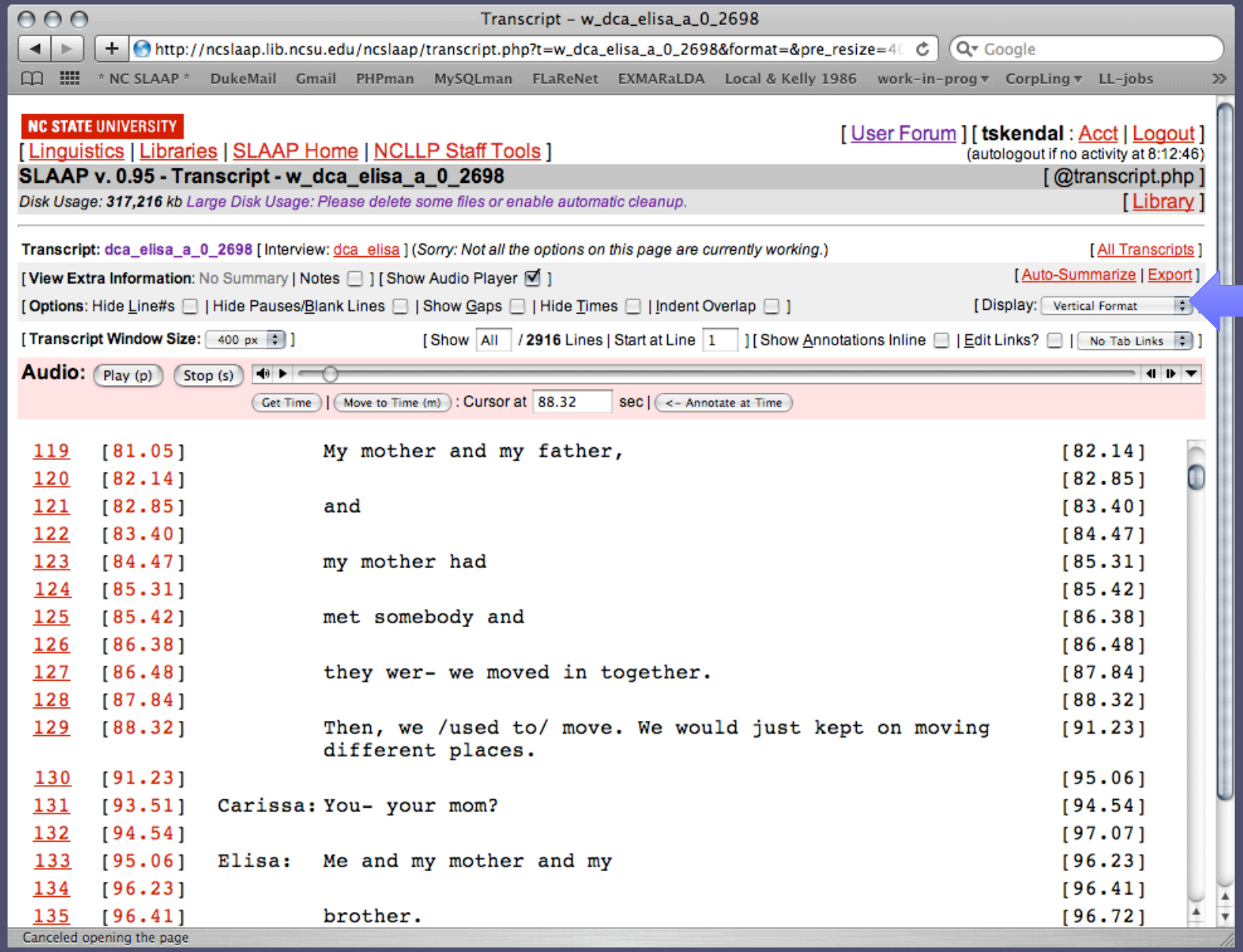


All records for a particular project – here “DC Adolescents Project”

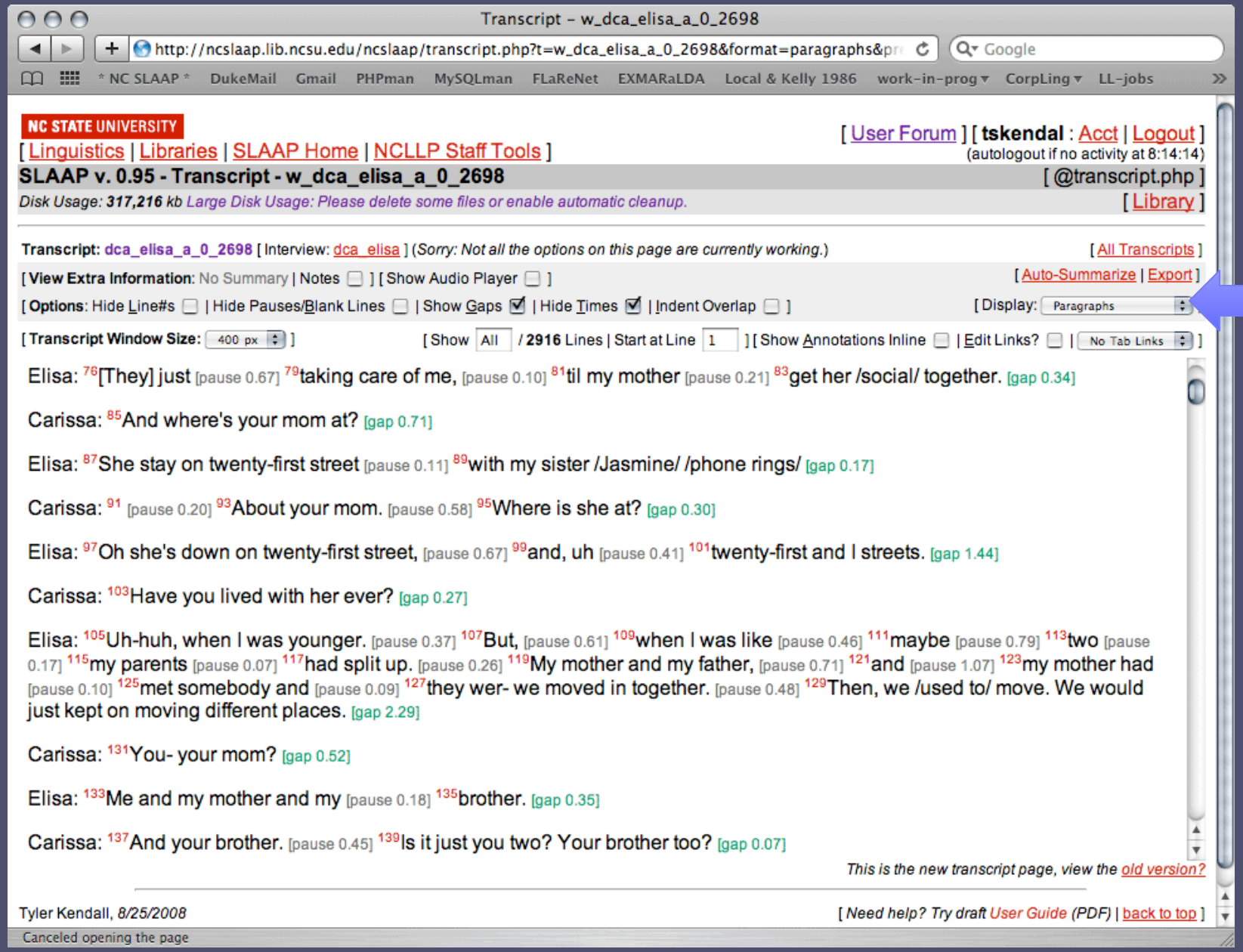


Transcript	Project	Interview/Media	Speaker(s)	Inter-viewer(s)	Length [sort]	Time Range (s) (% of Media File)	Num Lines	Special
dca_alayna_a_0_4245	DC Adolescents Project	dca_alayn / dca_alayna_a	Alayna	CFR	4245 sec (70.75 min)	0 - 4245 (100%)	4665	summarize stats|export
dca_asia_a_0_3358	DC Adolescents Project	dca_asia / dca_asia_a	Asia	CFR	3358 sec (55.97 min)	0 - 3358 (100%)	3388	summarize stats|export
dca_asia_b_0_655	DC Adolescents Project	dca_asia / dca_asia_b	Asia	CFR	655 sec (10.91 min)	0 - 655 (100%)	745	summarize stats|export
dca_calandra_a_0_2500	DC Adolescents Project	dca_calan / dca_calandra_a	Calandra	CFR	2500 sec (41.67 min)	0 - 2500 (100%)	2614	summarize stats|export
dca_edwin_a_0_3416	DC Adolescents Project	dca_edwin / dca_edwin_a	Edwin	CFR	3416 sec (56.93 min)	0 - 3416 (100%)	3738	summarize stats|export
dca_edwin_b_0_336	DC Adolescents Project	dca_edwin / dca_edwin_b	Edwin	CFR	336 sec (5.61 min)	0 - 336 (100%)	325	summarize stats|export
dca_elisa_a_0_2698	DC Adolescents Project	dca_elisa / dca_elisa_a	Elisa	CFR	2698 sec (44.96 min)	0 - 2698 (100%)	2916	summarize stats|export
dca_grace_a_0_3401	DC Adolescents Project	dca_grace / dca_grace_a	Grace	CFR	3401 sec (56.68 min)	0 - 3401 (100%)	3467	summarize stats|export
dca_grace_b_0_461	DC Adolescents Project	dca_grace / dca_grace_b	Grace	CFR	461 sec (7.68 min)	0 - 461 (100%)	488	summarize stats|export
dca_keisha_a_0_4007	DC Adolescents Project	dca_keish / dca_keisha_a	Keisha	CFR	4007 sec (66.78 min)	0 - 4007 (100%)	4473	summarize stats|export
dca_latania_a_0_3610	DC Adolescents Project	dca_latan / dca_latania_a	Latania	CFR	3610 sec (60.16 min)	0 - 3610 (100%)	4585	summarize stats|export
dca_shantell_a_0_3022	DC Adolescents Project	dca_shant / dca_shantell_a	Shantell	CFR	3022 sec (50.37 min)	0 - 3022 (100%)	3484	summarize stats|export
dca_shantell_b_0_1861	DC Adolescents Project	dca_shant / dca_shantell_b	Shantell	CFR	1861 sec (31.02 min)	0 - 1861 (100%)	1905	summarize stats|export
dca_shawna_a_0_2927	DC Adolescents Project	dca_shawn / dca_shawna_a	Shawna	CFR	2927 sec (48.79 min)	0 - 2927 (100%)	3082	summarize stats|export
dca_shirlisa_a_0_3434	DC Adolescents Project	dca_shirl / dca_shirlisa_a	Shirlisa	CFR	3434 sec (57.24 min)	0 - 3434 (100%)	3609	summarize stats|export

All transcripts for a particular project – here "DC Adolescents Project"

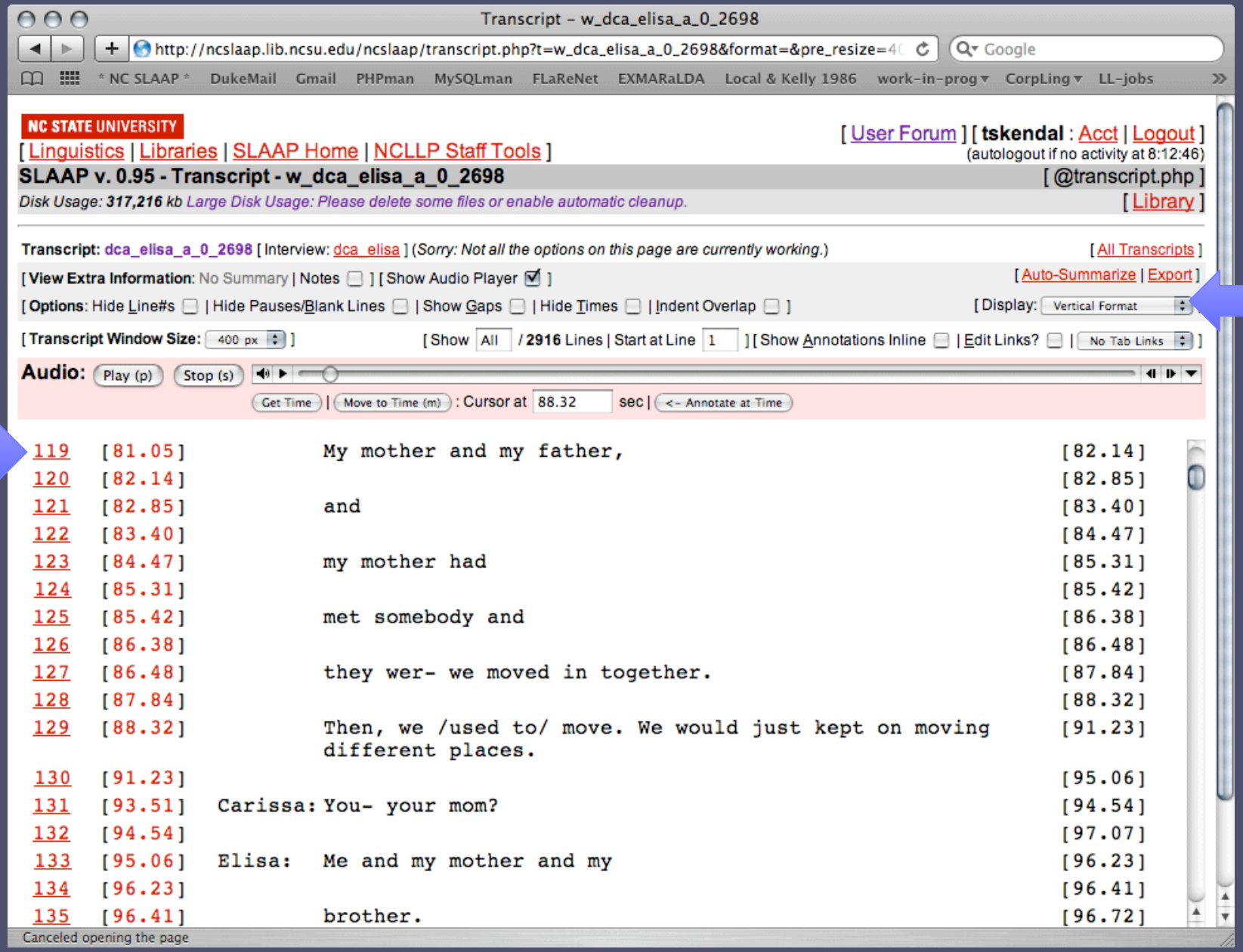


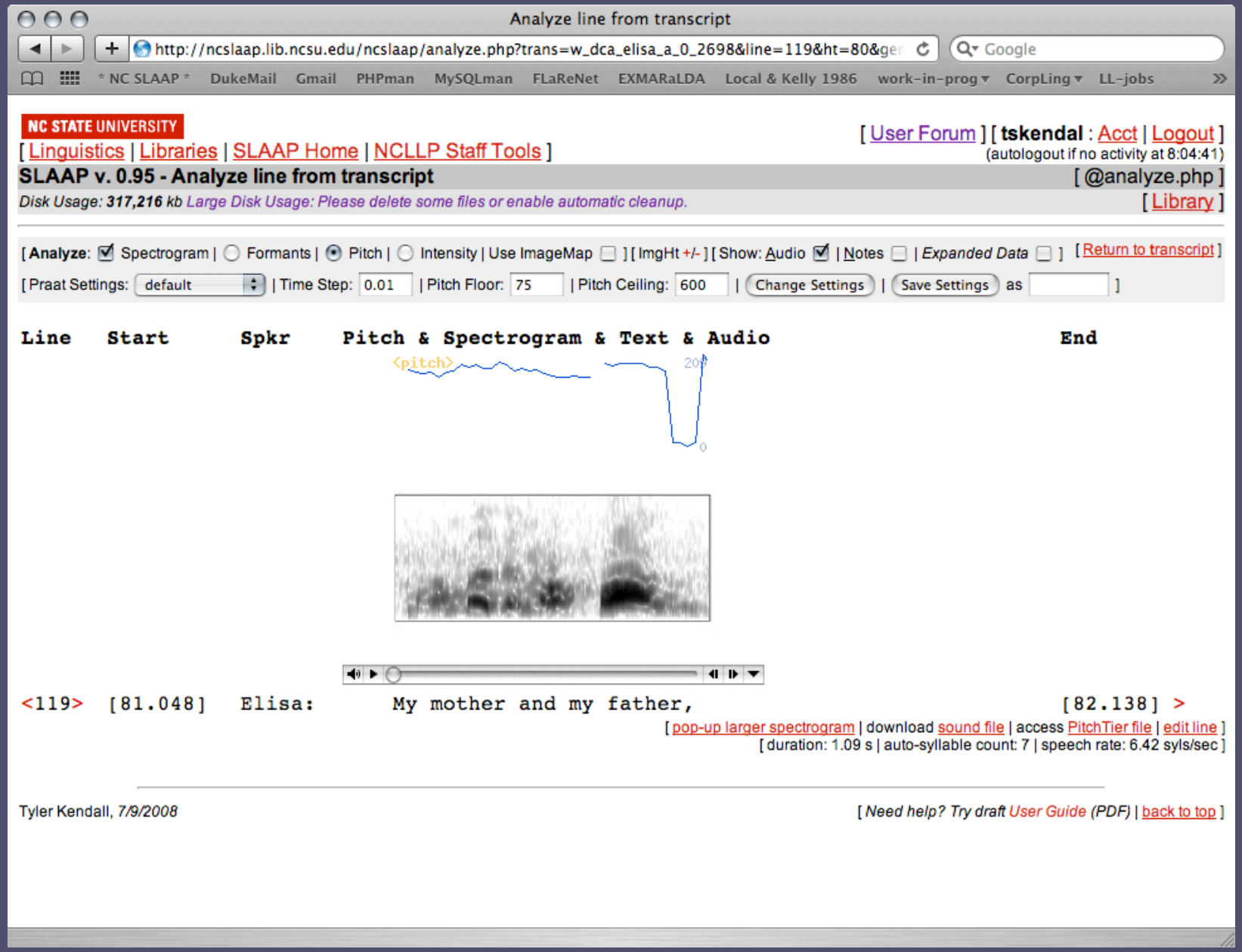
Multiple transcript views (w/ or w/out audio, etc.) for “Elisa” interview



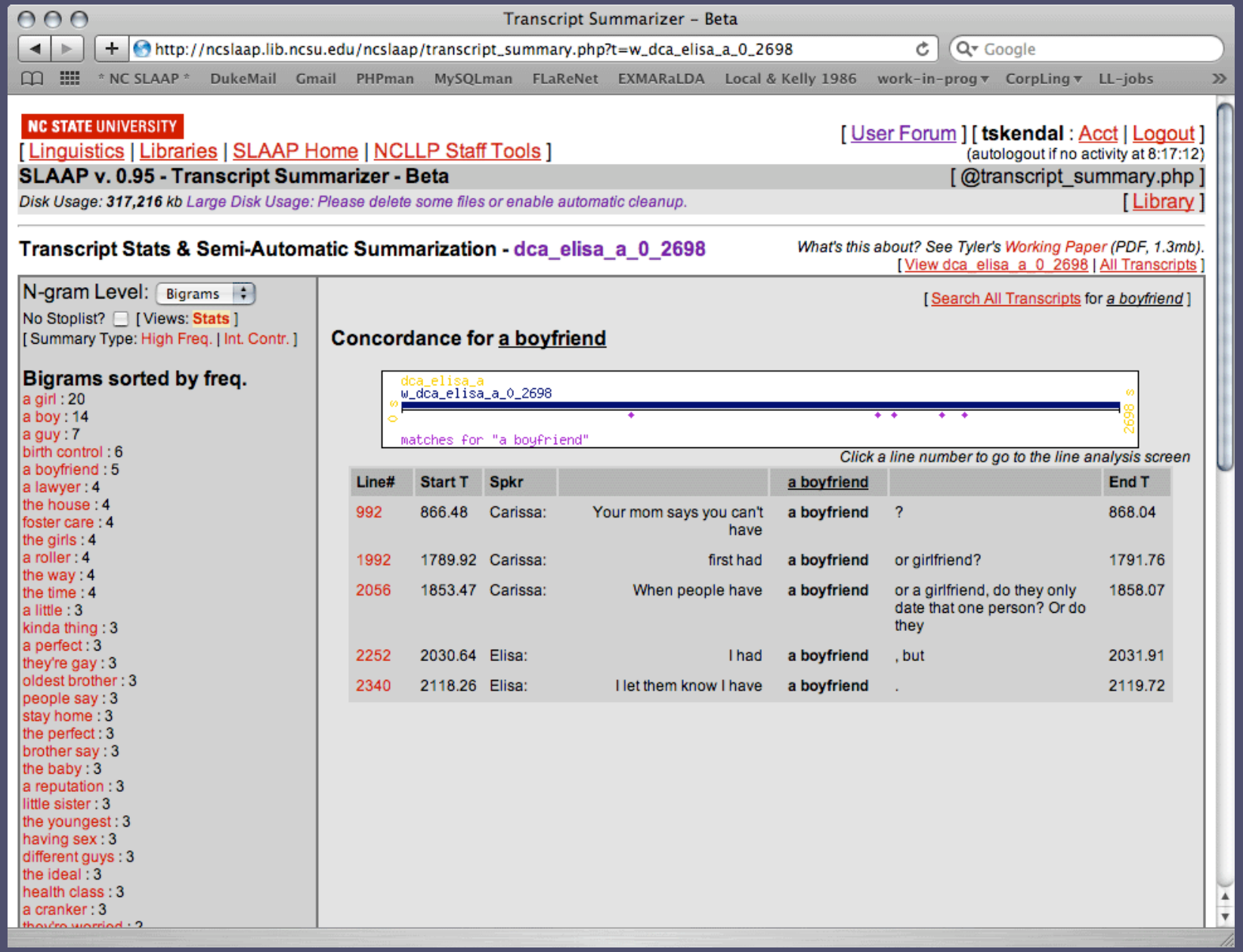
Multiple transcript views (here, “paragraphs”) for “Elisa” interview

Multiple transcript views (here “graphicalization”) for “Elisa” interview

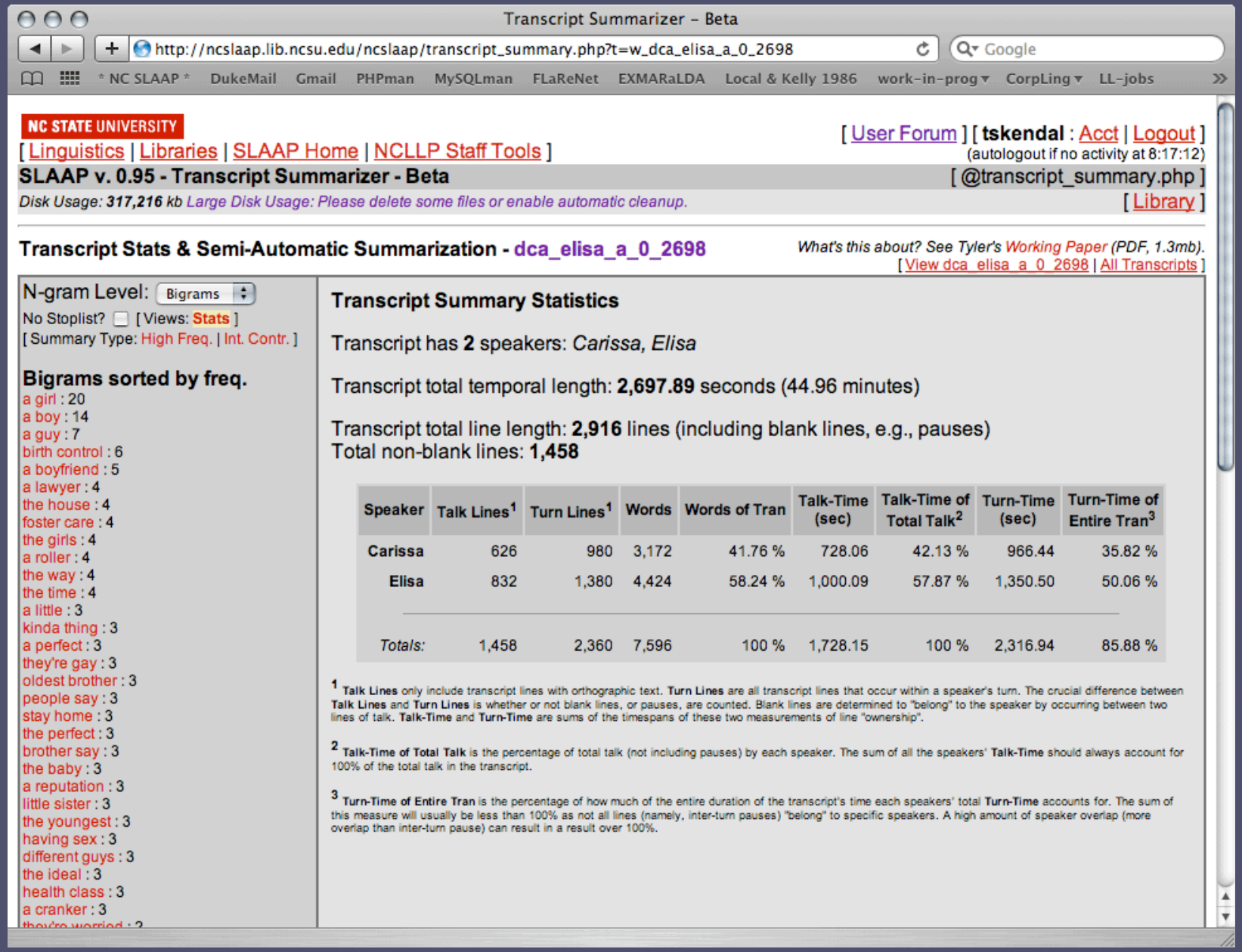

Transcript - w_dca_elisa_a_0_2698
http://ncslaap.lib.ncsu.edu/ncslaap/transcript.php?t=w_dca_elisa_a_0_2698&format=&pre_resize=
Google
NC STATE UNIVERSITY
[Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools]
[User Forum] [tskendal : Acct | Logout]
(autologout if no activity at 8:12:46)
SLAAP v. 0.95 - Transcript - w_dca_elisa_a_0_2698
[@transcript.php]
Disk Usage: 317,216 kb Large Disk Usage: Please delete some files or enable automatic cleanup.
[Library]
Transcript: dca_elisa_a_0_2698 [Interview: dca_elisa] (Sorry: Not all the options on this page are currently working.)
[All Transcripts]
[View Extra Information: No Summary | Notes] [Show Audio Player]
[Auto-Summarize | Export]
[Options: Hide Line#s | Hide Pauses/Blank Lines | Show Gaps | Hide Times | Indent Overlap]
[Display: Vertical Format
[Transcript Window Size: 400 px]
[Show All / 2916 Lines | Start at Line 1] [Show Annotations Inline | Edit Links? | No Tab Links]
Audio: Play (p) Stop (s)
Get Time | Move to Time (m) : Cursor at 88.32 sec | <- Annotate at Time
119 [81.05] My mother and my father, [82.14]
120 [82.14] [82.85]
121 [82.85] and [83.40]
122 [83.40] [84.47]
123 [84.47] my mother had [85.31]
124 [85.31] [85.42]
125 [85.42] met somebody and [86.38]
126 [86.38] [86.48]
127 [86.48] they wer- we moved in together. [87.84]
128 [87.84] [88.32]
129 [88.32] Then, we /used to/ move. We would just kept on moving different places. [91.23]
130 [91.23] [95.06]
131 [93.51] Carissa: You- your mom? [94.54]
132 [94.54] [97.07]
133 [95.06] Elisa: Me and my mother and my [96.23]
134 [96.23] [96.41]
135 [96.41] brother. [96.72]
Canceled opening the page



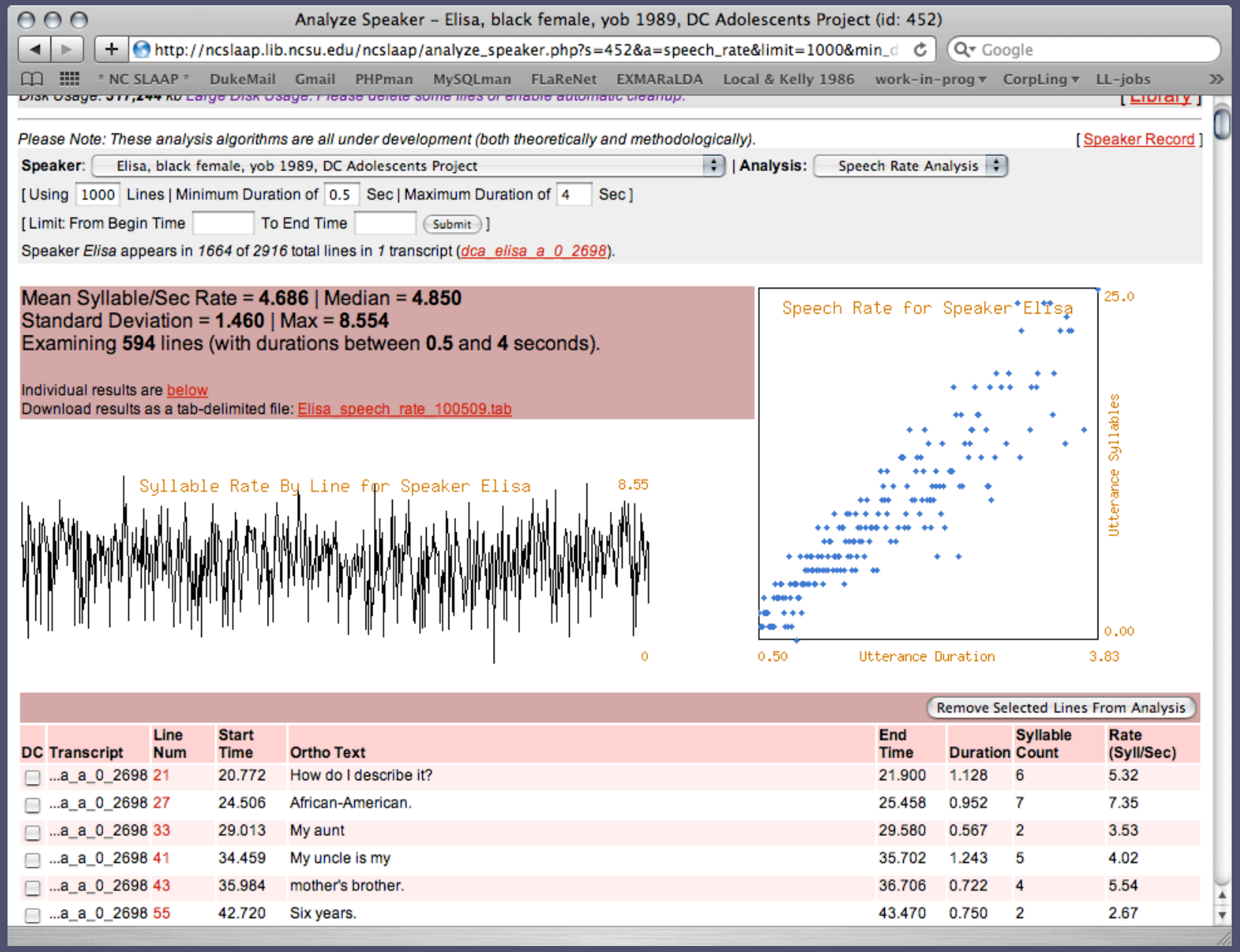
Can “analyze” specific lines; extract acoustic data on-the-fly



Explore the transcripts; frequent bigrams KWIC for “Elisa” interview



Explore the transcripts; summary statistics for "Elisa" interview



Analyze the transcripts; Speech rate for “Elisa” in her interview

SLAAP's database structure

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

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

Start Time	Speaker Reference	Simple Orthographic Representation	End Time	Add'l (meta)data & mark up

Adding a Transcript to the Archive

"Raw" Time-Stamped Transcript

Typically Praat TextGrid-based

SLAAP upload webpage*

Processing…

(e.g., linked to media file, speakers associated with databased entities)

Resides in SLAAP's database

SLAAP can process in a variety of ways

Can script with Praat or R or …

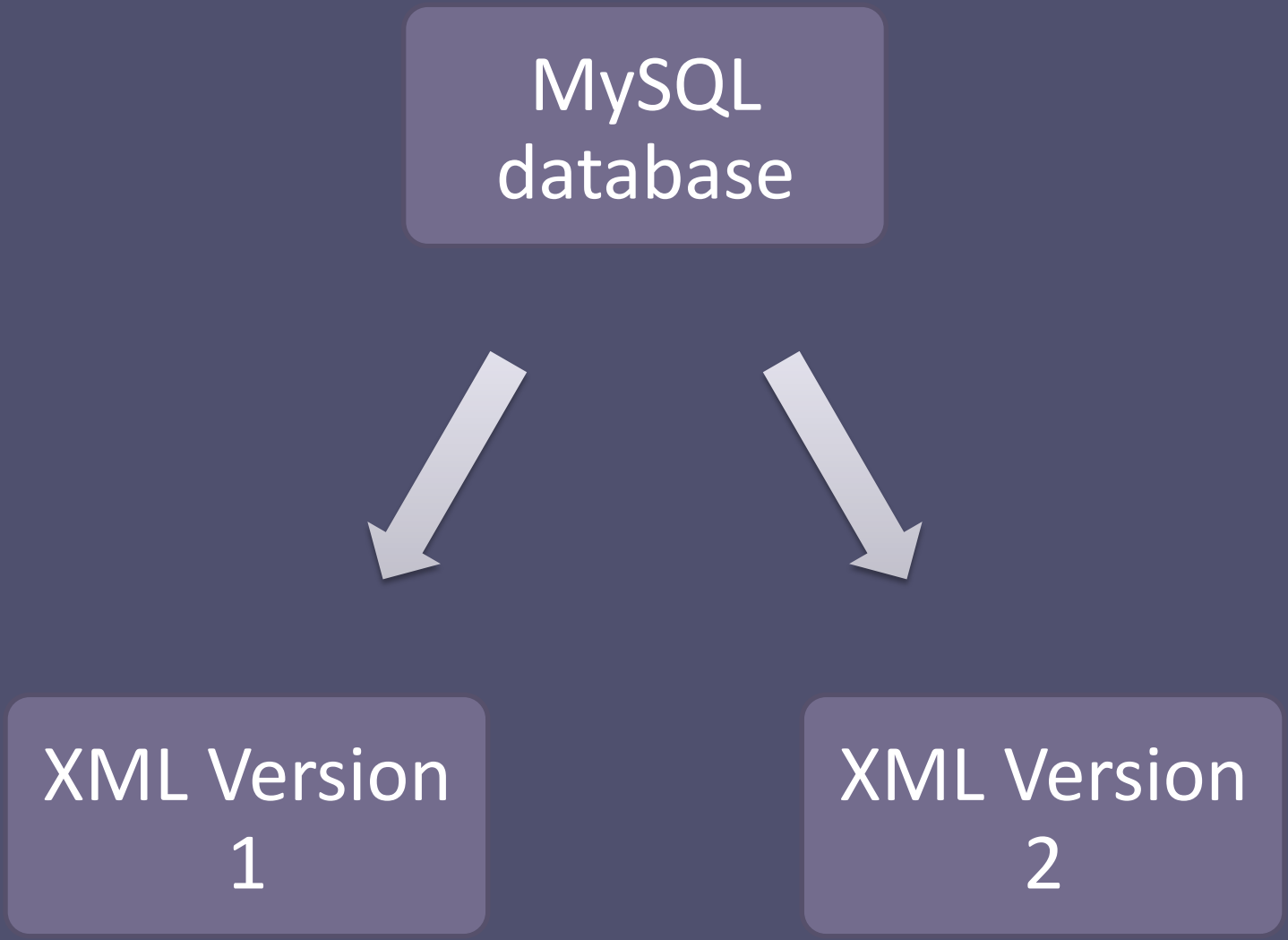
Can manipulate directly in MySQL

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

Etc.

* For non-SLAAP users, there's a public tool to convert TextGrids to plain text on the website

Storage in DB precedes XML



SLAAP itself works by…

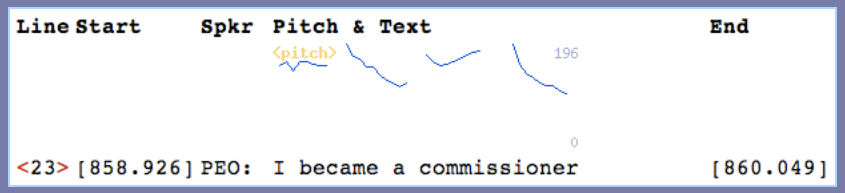
- All web pages are generated by PHP scripts
- Acoustic analysis features work by:
  - Extracting relevant information from the database (timestamps, media file location, etc.)
  - Sending this information to Praat via customized Praat scripts
  - Reading Praat output (and optionally post-processing,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

- E.g., extracting pitch info:

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

Praat

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



The value of relational databases

- This is not to say that SLAAP is possible because of its use of relational databases
- Or that it would be impossible with an XML backend

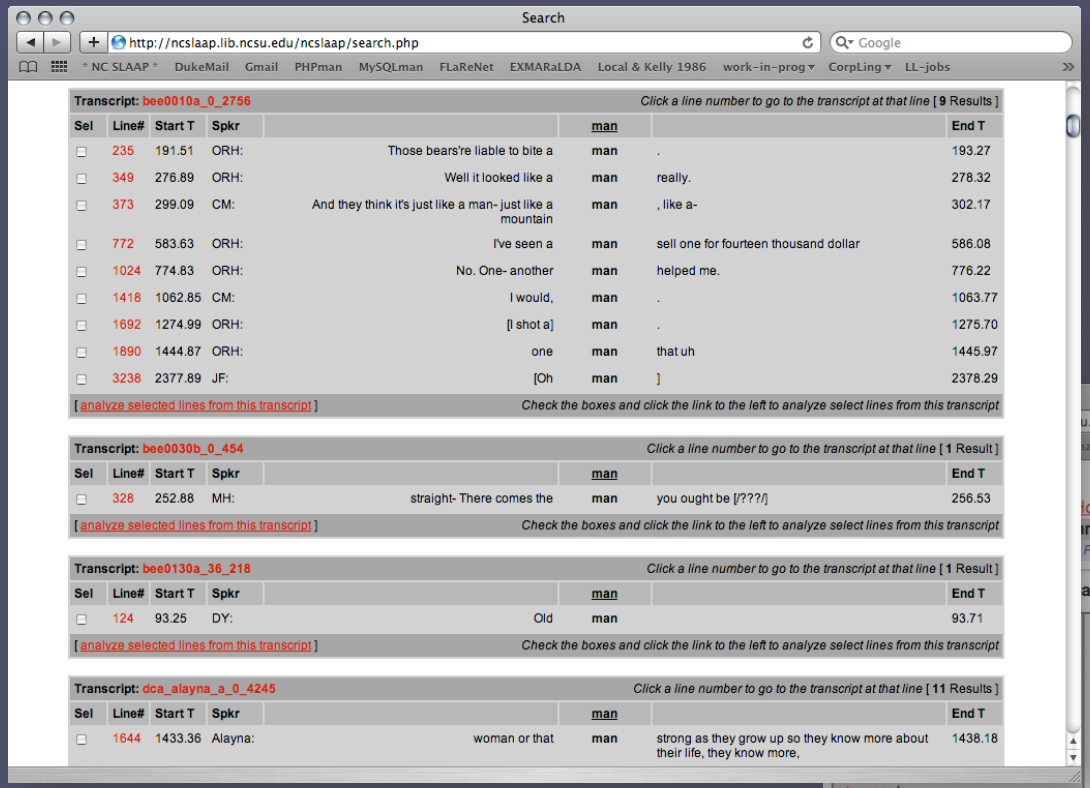
- However, the benefits of the database model seem clear in SLAAP
  - E.g., the ability to dynamically generate multiple versions of a “transcript”, both text-based and graphical is straightforwardly implemented with the DB backend. It’s harder (for me) to envision a simple way to do this in XML…

A broad question

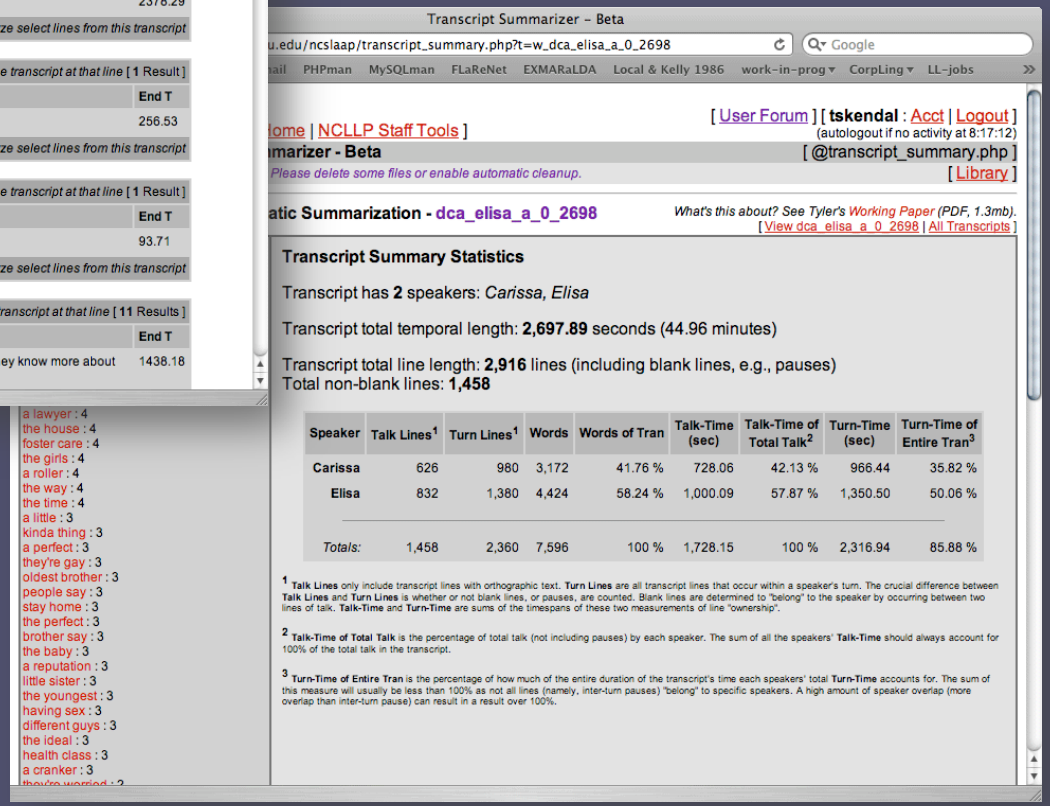
- Can we envision public corpora stored entirely in relational databases?
  - Accessed remotely through SQL, or a web-based (e.g., PHP) interface, or scripts in R, etc…
    - Variable interfaces, a separation of form and content (~XML)

- Corpus developers could grant free and open access to their materials, without "giving" away the actual corpus documents.
  - ~ Davies' excellent resources (http://corpus.byu.edu/)
    - ~ http://corpus.byu.edu/copyright.asp

E.g.,

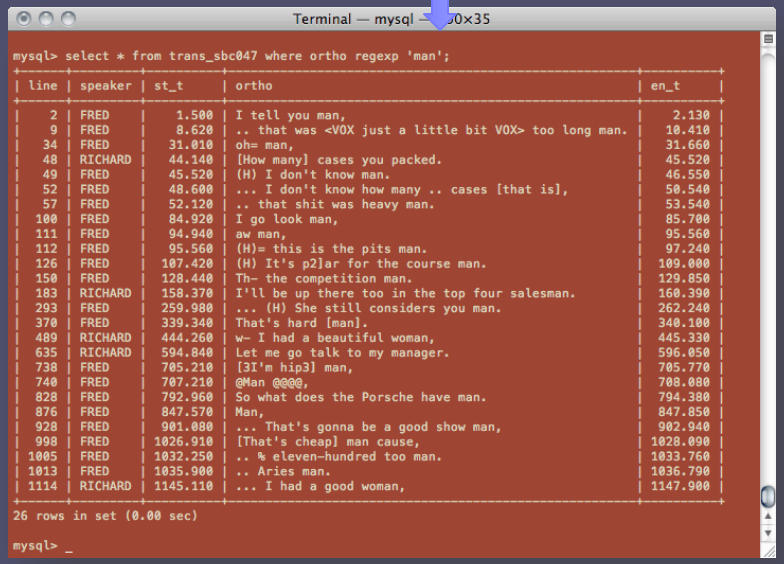


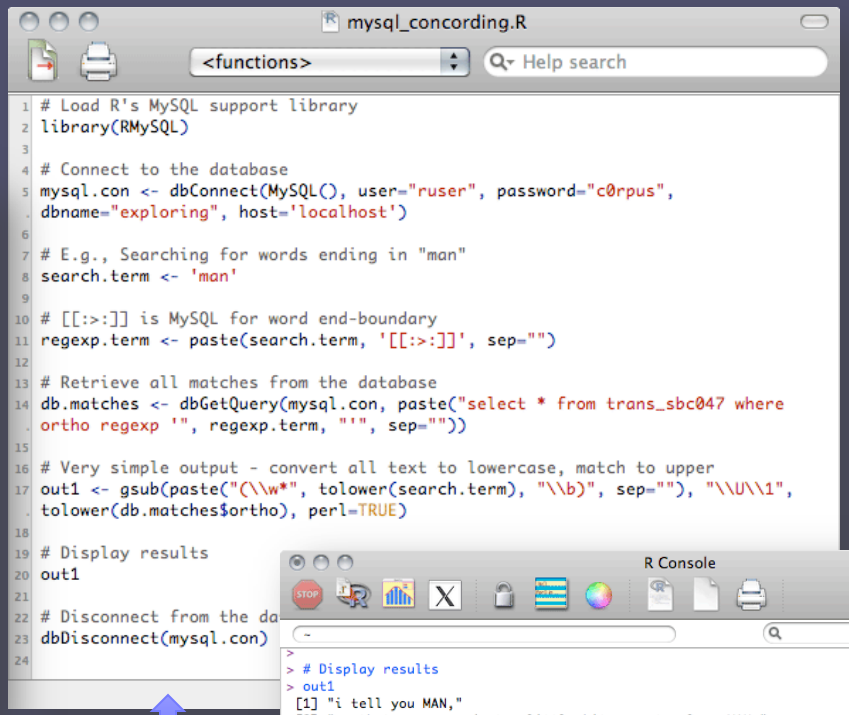
Can create SLAAP-like custom interfaces

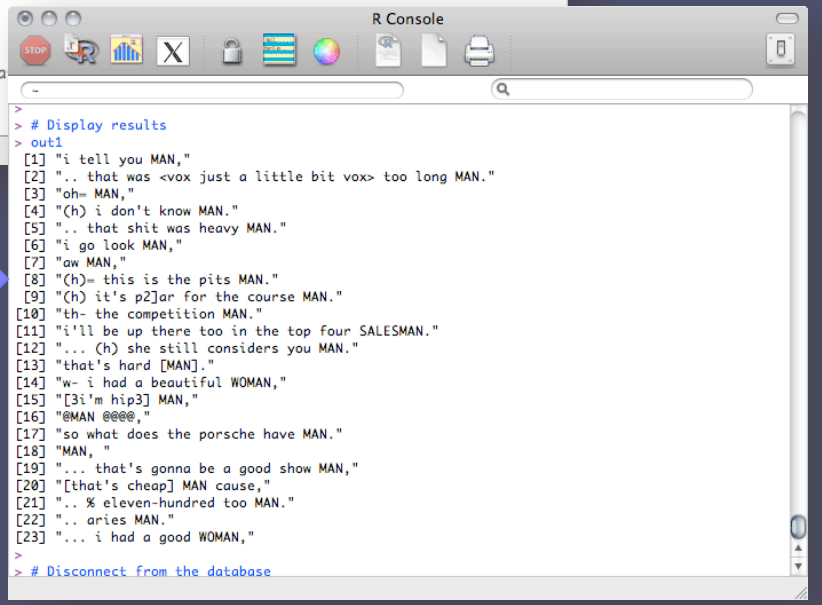


E.g.,

But could also access directly from remote MySQL clients







Or from R using the RMySQL library

Example from Santa Barbara Corpus (Du Bois et al. 2005) stored in MySQL DB

Thank you

- **Reactions / Thoughts?**

- More on SLAAP: http://ncslaap.lib.ncsu.edu/ (Kendall 2007, 2008)

Thanks to support from the North Carolina State University Libraries, the North Carolina Language and Life Project, and the William C. Friday Endowment at NC State University.

References

Bird, Steven and Mark Liberman. 2001. A formal framework of linguistic annotation. *Speech Communication* 33.1-2: 23-60.

Codd, Edgar F. 1970. A relational model of data for large shared data banks. *Communications of the ACM* 13.6: 377-387.

Davies, Mark. 2005. The advantage of using relational database for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics* 10.3: 307-334.

Du Bois, John W., and Englebretson, Robert. 2005. *Santa Barbara Corpus of Spoken American English*. Philadelphia: Linguistic Data Consortium.

Gries, Stefan Th. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York/ London: Routledge.

Kendall, Tyler. 2007. The North Carolina Sociolinguistic Archive and Analysis Project: Empowering the Sociolinguistic Archive. *Penn Working Papers in Linguistics* 13.2: Philadelphia: University of Pennsylvania. 15-26.

---. 2008. On the history and future of sociolinguistic data management. *Language and Linguistics Compass* 2.2: 332-351.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. New York: Routledge.

Poplack, Shana. 2007. Foreward. In J. Beal, K. Corrigan, & H. Moisl (eds), *Creating and Digitizing Language Corpora*. Volume 1: Synchronic Databases. New York/Basingstoke, Hampshire: Palgrave-Macmillan. ix-xiii.

Simons, Gary, William Lewis, Scott Farrar, D. Terence Langendoen, Brian Fitzsimons, and Hector Gonzalez. 2004. The semantics of markup: Mapping legacy markup schemas to a common semantics. *Proceedings of the 4th Workshop on NLP and XML (NLPXML-2004): Held in cooperation with ACL-04.* Barcelona, Spain. 25-32.