

# 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/aac12009-reldb.pdf>



AACL 2009, University of Alberta, Oct. 10<sup>th</sup> 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/>

**1** main library view

**2** full record view

**3** listen and annotate

**4** download and extract

**5** variable tabulation tool

**6** transcript features

**7** speaker-pitch analysis

Interview	Site	Speaker(s)	Interview Info	Media	Transcripts
prv007a	Princville PEO	black female, born 1964	Date: 09/26/2003 Interviewer(s): RR, DG Contains: sociolinguistic interview	prv007aa [Listen] [Download] prv007ab [Listen] [Download]	prv007aa_1980_2090 prv007aa_...
prv007b	Princville PEO	black female, born 1964	Date: 09/26/2003 Interviewer(s): RR, DG Contains: car tour of town	prv007ba [Listen] [Download] prv007bb [Listen] [Download]	
prv0110a	Princville SK	black male, age 55	Date: 10/03/2003 Interviewer(s): RR, DG Contains: sociolinguistic interview, ?	prv0110a [Listen] [Download] prv0110b [Listen] [Download]	prv0110a_...
prv0111f	Princville SK	black male, age 55	Date: 02/21/2005 Interviewer(s): RJ Contains: radio interview	pvlv0111f [Listen] [Download] pvlv0112f [Listen] [Download] pvlv0113f [Listen] [Download] pvlv0114f [Listen] [Download] pvlv0115f [Listen] [Download] pvlv0116f [Listen] [Download] pvlv0117f [Listen] [Download]	pvlv0115f_5 pvlv0115f_8
prv021v	Princville PEO	black female, born 1964	Date: 02/18/2005 Interviewer(s): DG	prv021v [Listen] [Download] prv021v [Listen] [Download] prv021v [Listen] [Download]	pvlv021v_5

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

Main Library Access

http://ncslaap.lib.ncsu.edu/ncslaap/library.php?startat=0&view=long&browse\_project=11&at\_a\_time

NC STATE UNIVERSITY

[Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools]

[ User Forum ] [ tskendal : Acct | Logout ] (autologout if no activity at 8:25:12)

SLAAP v. 0.95 - Main Library Access [ @library.php ]

Disk Usage: 317,532 kb Large Disk Usage: Please delete some files or enable automatic cleanup.

[ Users | Tabulation Summary | Speaker Analysis | Lexical Analysis | Manage Soundfiles and Metadata ]

SLAAP Archive: Browse | Filter | Projects | Speakers | Transcripts [ All Archive Search ] 1481 total records in SLAAP

[ View: Long | Short ] [ Show Project Info? ] [ 15 records at a time ]

[ Project: DC Adolescents Project ] Showing 15 (of 19 records) [ Page: 1, 2 | >> ]

Interview	Project	Speaker(s)	Interview Info	Media	Transcripts
<a href="#">dca_alayn</a> [ Full Record ]	DC Adolescents Project	Alayna Black Female, Born 1989 Locality: Washington, D.C.	Date: 07/01/2003 Interviewer(s): CFR Language(s): English Contains: sociological interview ?	dca_alayna_a [ Listen   Download ]	<a href="#">dca_alayna_a_0_4245</a>
<a href="#">dca_asia</a> [ Full Record ]	DC Adolescents Project	Asia Black Female, Born 1991 Locality: Washington, D.C.	Date: 07/01/2003 Interviewer(s): CFR Language(s): English Contains: sociological interview ?	dca_asia_a [ Listen   Download ] dca_asia_b [ Listen   Download ] **	<a href="#">dca_asia_a_0_3358</a> <a href="#">dca_asia_b_0_655</a>
<a href="#">dca_calan</a> [ Full Record ]	DC Adolescents Project	Calandra Black Female, Born 1990 Locality: Washington, D.C.	Date: 07/01/2003 Interviewer(s): CFR Language(s): English Contains: sociological interview ?	dca_calandra_a [ Listen   Download ]	<a href="#">dca_calandra_a_0_2500</a>
<a href="#">dca_dariu</a> [ Full Record ]	DC Adolescents Project	Darius Black Male, Born 1990 Locality: Washington, D.C.	Date: 07/01/2003 Interviewer(s): CFR Language(s): English Contains: sociological interview ?	dca_darius_a [ Listen   Download ]	
<a href="#">dca_edwin</a> [ Full Record ]	DC Adolescents Project	Edwin Black Male, Born 1990 Locality: Washington, D.C.	Date: 07/01/2003 Interviewer(s): CFR Language(s): English Contains: sociological interview	dca_edwin_a [ Listen   Download ] dca_edwin_b [ Listen   Download ]	<a href="#">dca_edwin_a_0_3416</a> <a href="#">dca_edwin_b_0_336</a>
<a href="#">dca_elisa</a> [ Full Record ]	DC Adolescents Project	Elisa Black Female, Born 1989 Locality: Washington, D.C.	Date: 07/01/2003 Interviewer(s): CFR Language(s): English Contains: sociological interview	dca_elisa_a [ Listen   Download ]	<a href="#">dca_elisa_a_0_2698</a>

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

Transcript Summary

http://ncslaap.lib.ncsu.edu/ncslaap/transcripts.php?proj\_id=11

NC STATE UNIVERSITY

[ Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools ]

SLAAP v. 0.95 - Transcript Summary

Disk Usage: 317,532 kb Large Disk Usage: Please delete some files or enable automatic cleanup.

[ User Forum ] [ tskendal : Acct | Logout ] (autologout if no activity at 8:24:46)

[ @transcripts.php ] [ Library ]

[ Search Transcripts ] Show Transcripts for: DC Adolescents Project

18 transcripts are currently available to you for DC Adolescents Project, covering 44,575 seconds (about 12.38 hours) of audio (SLAAP has 145 total transcripts).

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”

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=400

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) [ 88.32 SEC ]

<a href="#">119</a>	[ 81.05 ]	My mother and my father,	[ 82.14 ]
<a href="#">120</a>	[ 82.14 ]		[ 82.85 ]
<a href="#">121</a>	[ 82.85 ]	and	[ 83.40 ]
<a href="#">122</a>	[ 83.40 ]		[ 84.47 ]
<a href="#">123</a>	[ 84.47 ]	my mother had	[ 85.31 ]
<a href="#">124</a>	[ 85.31 ]		[ 85.42 ]
<a href="#">125</a>	[ 85.42 ]	met somebody and	[ 86.38 ]
<a href="#">126</a>	[ 86.38 ]		[ 86.48 ]
<a href="#">127</a>	[ 86.48 ]	they wer- we moved in together.	[ 87.84 ]
<a href="#">128</a>	[ 87.84 ]		[ 88.32 ]
<a href="#">129</a>	[ 88.32 ]	Then, we /used to/ move. We would just kept on moving different places.	[ 91.23 ]
<a href="#">130</a>	[ 91.23 ]		[ 95.06 ]
<a href="#">131</a>	[ 93.51 ]	Carissa: You- your mom?	[ 94.54 ]
<a href="#">132</a>	[ 94.54 ]		[ 97.07 ]
<a href="#">133</a>	[ 95.06 ]	Elisa: Me and my mother and my	[ 96.23 ]
<a href="#">134</a>	[ 96.23 ]		[ 96.41 ]
<a href="#">135</a>	[ 96.41 ]	brother.	[ 96.72 ]

Canceled opening the page

## Multiple transcript views (w/ or w/out audio, etc.) 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=paragraphs&pr

NC STATE UNIVERSITY

[Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools]

SLAAP v. 0.95 - Transcript - w\_dca\_elisa\_a\_0\_2698

Disk Usage: 317,216 kb Large Disk Usage: Please delete some files or enable automatic cleanup.

Transcript: dca\_elisa\_a\_0\_2698 [Interview: dca\_elisa] (Sorry: Not all the options on this page are currently working.)

[View Extra Information: No Summary | Notes ] [Show Audio Player

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

[Transcript Window Size: 400 px] [Show All / 2916 Lines | Start at Line 1] [Show Annotations Inline  | Edit Links?  | No Tab Links

Elisa: <sup>76</sup>[They] just [pause 0.67] <sup>79</sup>taking care of me, [pause 0.10] <sup>81</sup>til my mother [pause 0.21] <sup>83</sup>get her /social/ together. [gap 0.34]

Carissa: <sup>85</sup>And where's your mom at? [gap 0.71]

Elisa: <sup>87</sup>She stay on twenty-first street [pause 0.11] <sup>89</sup>with my sister /Jasmine/ /phone rings/ [gap 0.17]

Carissa: <sup>91</sup> [pause 0.20] <sup>93</sup>About your mom. [pause 0.58] <sup>95</sup>Where is she at? [gap 0.30]

Elisa: <sup>97</sup>Oh she's down on twenty-first street, [pause 0.67] <sup>99</sup>and, uh [pause 0.41] <sup>101</sup>twenty-first and I streets. [gap 1.44]

Carissa: <sup>103</sup>Have you lived with her ever? [gap 0.27]

Elisa: <sup>105</sup>Uh-huh, when I was younger. [pause 0.37] <sup>107</sup>But, [pause 0.61] <sup>109</sup>when I was like [pause 0.46] <sup>111</sup>maybe [pause 0.79] <sup>113</sup>two [pause 0.17] <sup>115</sup>my parents [pause 0.07] <sup>117</sup>had split up. [pause 0.26] <sup>119</sup>My mother and my father, [pause 0.71] <sup>121</sup>and [pause 1.07] <sup>123</sup>my mother had [pause 0.10] <sup>125</sup>met somebody and [pause 0.09] <sup>127</sup>they wer- we moved in together. [pause 0.48] <sup>129</sup>Then, we /used to/ move. We would just kept on moving different places. [gap 2.29]

Carissa: <sup>131</sup>You- your mom? [gap 0.52]

Elisa: <sup>133</sup>Me and my mother and my [pause 0.18] <sup>135</sup>brother. [gap 0.35]

Carissa: <sup>137</sup>And your brother. [pause 0.45] <sup>139</sup>Is it just you two? Your brother too? [gap 0.07]

This is the new transcript page, view the [old version?](#)

Tyler Kendall, 8/25/2008

[Need help? Try draft [User Guide \(PDF\)](#) | [back to top](#)]

Canceled opening the page



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

Transcript - w\_dca\_elisa\_a\_0\_2698 - Time Line

http://ncslaap.lib.ncsu.edu/ncslaap/tran\_timeline.php?t=w\_dca\_elisa\_a\_0\_2698&start\_line=0&num\_li

NC STATE UNIVERSITY [ User Forum ] [ tskendal : Acct | Logout ]  
 (autologout if no activity at 8:16:19)

[ Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools ] [ @tran\_timeline.php ]

SLAAP v. 0.95 - Transcript - w\_dca\_elisa\_a\_0\_2698 - Time Line [ Library ]

Disk Usage: 317,216 kb Large Disk Usage: Please delete some files or enable automatic cleanup.

Transcript: dca\_elisa\_a\_0\_2698 [ Interview: dca\_elisa ] [ Transcript Summary ]

[ Settings: Resolution 6 pixels per second | Use ImageMap ] [ Display: Graphicalization ]

[ Show Annotations (if avail.) ] [ Show Tabulations (if avail.) ] [ All Tabs ] [ Show Audio: ]

Graphicalization of transcript w\_dca\_elisa\_a\_0\_2698

Carissa: [ graphical representation ]

Elisa: [ graphical representation ]

darker depicts higher speech rates (syl/sec): >10 >8 >6 >4 >2 >0  
 colors of tabs indicate vernacularity: most vernacular most standard

Tyler Kendall, 5/27/2008 [ Need help? Try draft User Guide (PDF) | back to top ]

## 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=400

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) [ 88.32 SEC ]

<a href="#">119</a>	[ 81.05 ]	My mother and my father,	[ 82.14 ]
<a href="#">120</a>	[ 82.14 ]		[ 82.85 ]
<a href="#">121</a>	[ 82.85 ]	and	[ 83.40 ]
<a href="#">122</a>	[ 83.40 ]		[ 84.47 ]
<a href="#">123</a>	[ 84.47 ]	my mother had	[ 85.31 ]
<a href="#">124</a>	[ 85.31 ]		[ 85.42 ]
<a href="#">125</a>	[ 85.42 ]	met somebody and	[ 86.38 ]
<a href="#">126</a>	[ 86.38 ]		[ 86.48 ]
<a href="#">127</a>	[ 86.48 ]	they wer- we moved in together.	[ 87.84 ]
<a href="#">128</a>	[ 87.84 ]		[ 88.32 ]
<a href="#">129</a>	[ 88.32 ]	Then, we /used to/ move. We would just kept on moving different places.	[ 91.23 ]
<a href="#">130</a>	[ 91.23 ]		[ 95.06 ]
<a href="#">131</a>	[ 93.51 ]	Carissa: You- your mom?	[ 94.54 ]
<a href="#">132</a>	[ 94.54 ]		[ 97.07 ]
<a href="#">133</a>	[ 95.06 ]	Elisa: Me and my mother and my	[ 96.23 ]
<a href="#">134</a>	[ 96.23 ]		[ 96.41 ]
<a href="#">135</a>	[ 96.41 ]	brother.	[ 96.72 ]

Canceled opening the page



Analyze line from transcript

http://ncslaap.lib.ncsu.edu/ncslaap/analyze.php?trans=w\_dca\_elisa\_a\_0\_2698&line=119&ht=80&ger

NC STATE UNIVERSITY


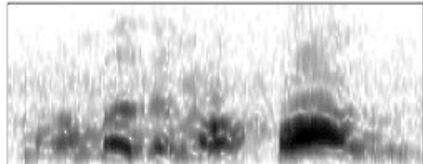
[ Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools ]

SLAAP v. 0.95 - Analyze line from transcript

Disk Usage: 317,216 kb Large Disk Usage: Please delete some files or enable automatic cleanup.

[ Analyze:  Spectrogram |  Formants |  Pitch |  Intensity | Use ImageMap  ] [ ImgHt +/- ] [ Show: Audio  | Notes  | Expanded Data  ] [ Return to transcript ]

[ Praat Settings: default | Time Step: 0.01 | Pitch Floor: 75 | Pitch Ceiling: 600 | Change Settings | Save Settings as ]

Line	Start	Spkr	Pitch & Spectrogram & Text & Audio	End
			 	
<119>	[81.048]	Elisa:	My mother and my father,	[82.138] >

[ pop-up larger spectrogram | download sound file | access PitchTier file | edit line ]  
 [ duration: 1.09 s | auto-syllable count: 7 | speech rate: 6.42 syls/sec ]

Tyler Kendall, 7/9/2008

[ Need help? Try draft User Guide (PDF) | back to top ]

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



Transcript Summarizer - Beta

http://ncslaap.lib.ncsu.edu/ncslaap/transcript\_summary.php?t=w\_dca\_elisa\_a\_0\_2698

NC STATE UNIVERSITY

[Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools]

SLAAP v. 0.95 - Transcript Summarizer - Beta

Disk Usage: 317,216 kb Large Disk Usage: Please delete some files or enable automatic cleanup.

[ User Forum ] [ tskendal : Acct | Logout ] (autologout if no activity at 8:17:12)

[ @transcript\_summary.php ] [ Library ]

Transcript Stats & Semi-Automatic Summarization - dca\_elisa\_a\_0\_2698

What's this about? See Tyler's Working Paper (PDF, 1.3mb). [ View dca\_elisa\_a\_0\_2698 | All Transcripts ]

N-gram Level: Bigrams

No Stoplist? [ Views: Stats ] [ Summary Type: High Freq. | Int. Contr. ]

Bigrams sorted by freq.

- a girl : 20
- a boy : 14
- a guy : 7
- birth control : 6
- a boyfriend : 5
- a lawyer : 4
- the house : 4
- foster care : 4
- the girls : 4
- a roller : 4
- the way : 4
- the time : 4
- a little : 3
- kinda thing : 3
- a perfect : 3
- they're gay : 3
- oldest brother : 3
- people say : 3
- stay home : 3
- the perfect : 3
- brother say : 3
- the baby : 3
- a reputation : 3
- little sister : 3
- the youngest : 3
- having sex : 3
- different guys : 3
- the ideal : 3
- health class : 3
- a cranker : 3
- they're worried : 2

Concordance for a boyfriend

Click a line number to go to the line analysis screen

Line#	Start T	Spkr		a boyfriend	End T
992	866.48	Carissa:	Your mom says you can't have	a boyfriend ?	868.04
1992	1789.92	Carissa:	first had	a boyfriend or girlfriend?	1791.76
2056	1853.47	Carissa:	When people have	a boyfriend or a girlfriend, do they only date that one person? Or do they	1858.07
2252	2030.64	Elisa:	I had	a boyfriend , but	2031.91
2340	2118.26	Elisa:	I let them know I have	a boyfriend .	2119.72

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

Transcript Summarizer - Beta

http://ncslaap.lib.ncsu.edu/ncslaap/transcript\_summary.php?t=w\_dca\_elisa\_a\_0\_2698

NC STATE UNIVERSITY

[ Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools ]

SLAAP v. 0.95 - Transcript Summarizer - Beta

Disk Usage: 317,216 kb Large Disk Usage: Please delete some files or enable automatic cleanup.

[ User Forum ] [ tskendal : Acct | Logout ] (autologout if no activity at 8:17:12)

[ @transcript\_summary.php ] [ Library ]

Transcript Stats & Semi-Automatic Summarization - dca\_elisa\_a\_0\_2698

What's this about? See Tyler's Working Paper (PDF, 1.3mb). [ View dca\_elisa\_a\_0\_2698 | All Transcripts ]

N-gram Level: Bigrams

No Stoplist? [ Views: Stats ]

[ Summary Type: High Freq. | Int. Contr. ]

**Bigrams sorted by freq.**

a girl : 20  
a boy : 14  
a guy : 7  
birth control : 6  
a boyfriend : 5  
a lawyer : 4  
the house : 4  
foster care : 4  
the girls : 4  
a roller : 4  
the way : 4  
the time : 4  
a little : 3  
kinda thing : 3  
a perfect : 3  
they're gay : 3  
oldest brother : 3  
people say : 3  
stay home : 3  
the perfect : 3  
brother say : 3  
the baby : 3  
a reputation : 3  
little sister : 3  
the youngest : 3  
having sex : 3  
different guys : 3  
the ideal : 3  
health class : 3  
a cranker : 3  
they're worried : 2

**Transcript Summary Statistics**

Transcript has 2 speakers: *Carissa, Elisa*

Transcript total temporal length: 2,697.89 seconds (44.96 minutes)

Transcript total line length: 2,916 lines (including blank lines, e.g., pauses)

Total non-blank lines: 1,458

Speaker	Talk Lines <sup>1</sup>	Turn Lines <sup>1</sup>	Words	Words of Tran	Talk-Time (sec)	Talk-Time of Total Talk <sup>2</sup>	Turn-Time (sec)	Turn-Time of Entire Tran <sup>3</sup>
Carissa	626	980	3,172	41.76 %	728.06	42.13 %	966.44	35.82 %
Elisa	832	1,380	4,424	58.24 %	1,000.09	57.87 %	1,350.50	50.06 %
Totals:	1,458	2,360	7,596	100 %	1,728.15	100 %	2,316.94	85.88 %

<sup>1</sup> Talk Lines only include transcript lines with orthographic text. Turn Lines are all transcript lines that occur within a speaker's turn. The crucial difference between Talk Lines and Turn Lines is whether or not blank lines, or pauses, are counted. Blank lines are determined to "belong" to the speaker by occurring between two lines of talk. Talk-Time and Turn-Time are sums of the timespans of these two measurements of line "ownership".

<sup>2</sup> Talk-Time of Total Talk is the percentage of total talk (not including pauses) by each speaker. The sum of all the speakers' Talk-Time should always account for 100% of the total talk in the transcript.

<sup>3</sup> Turn-Time of Entire Tran is the percentage of how much of the entire duration of the transcript's time each speakers' total Turn-Time accounts for. The sum of this measure will usually be less than 100% as not all lines (namely, inter-turn pauses) "belong" to specific speakers. A high amount of speaker overlap (more overlap than inter-turn pause) can result in a result over 100%.

## Explore the transcripts; summary statistics for "Elisa" interview

Analyze Speaker – Elisa, black female, yob 1989, DC Adolescents Project (id: 452)

http://ncslaap.lib.ncsu.edu/ncslaap/analyze\_speaker.php?s=452&a=speech\_rate&limit=1000&min\_d

Please Note: These analysis algorithms are all under development (both theoretically and methodologically). [Speaker Record]

Speaker: Elisa, black female, yob 1989, DC Adolescents Project | Analysis: Speech Rate Analysis

[Using 1000 Lines | Minimum Duration of 0.5 Sec | Maximum Duration of 4 Sec]

[Limit: From Begin Time To End Time Submit]

Speaker Elisa appears in 1664 of 2916 total lines in 1 transcript ([dca\\_elisa\\_a\\_0\\_2698](#)).

Mean Syllable/Sec Rate = 4.686 | Median = 4.850  
 Standard Deviation = 1.460 | Max = 8.554  
 Examining 594 lines (with durations between 0.5 and 4 seconds).

Individual results are [below](#)  
 Download results as a tab-delimited file: [Elisa\\_speech\\_rate\\_100509.tab](#)

Syllable Rate By Line for Speaker Elisa

Speech Rate for Speaker Elisa

Remove Selected Lines From Analysis

DC Transcript	Line Num	Start Time	Ortho Text	End Time	Duration	Syllable Count	Rate (Syll/Sec)
<input type="checkbox"/> ...a_a_0_2698	21	20.772	How do I describe it?	21.900	1.128	6	5.32
<input type="checkbox"/> ...a_a_0_2698	27	24.506	African-American.	25.458	0.952	7	7.35
<input type="checkbox"/> ...a_a_0_2698	33	29.013	My aunt	29.580	0.567	2	3.53
<input type="checkbox"/> ...a_a_0_2698	41	34.459	My uncle is my	35.702	1.243	5	4.02
<input type="checkbox"/> ...a_a_0_2698	43	35.984	mother's brother.	36.706	0.722	4	5.54
<input type="checkbox"/> ...a_a_0_2698	55	42.720	Six years.	43.470	0.750	2	2.67

# 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)

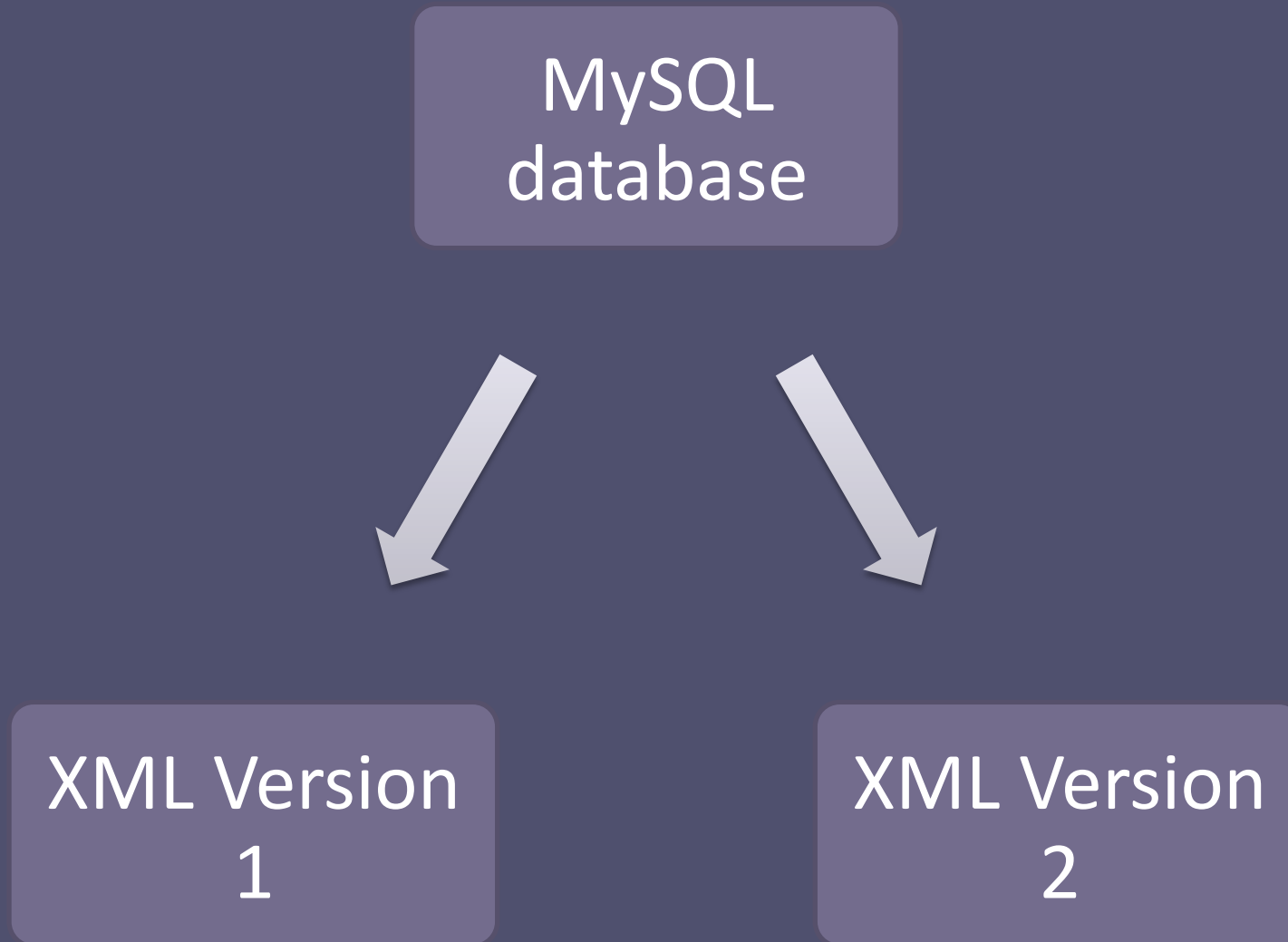


Etc.

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

Slides: <http://ncslaap.lib.ncsu.edu/pdfs/aac12009-reldb.pdf>

# 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.

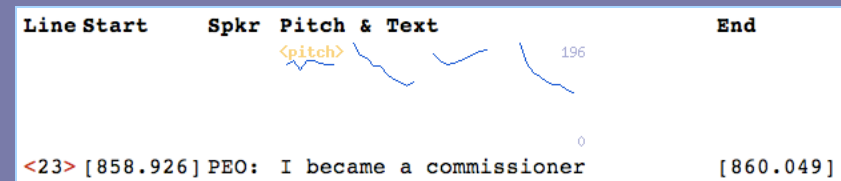
- E.g., extracting pitch info:

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.,

Search

http://ncslaap.lib.ncsu.edu/ncslaap/search.php

Transcript: **bee0010a\_0\_2756** Click a line number to go to the transcript at that line [ 9 Results ]

Sel	Line#	Start T	Spkr		man		End T
<input type="checkbox"/>	235	191.51	ORH:	Those bears're liable to bite a	man	.	193.27
<input type="checkbox"/>	349	276.89	ORH:	Well it looked like a	man	really.	278.32
<input type="checkbox"/>	373	299.09	CM:	And they think it's just like a man- just like a mountain	man	, like a-	302.17
<input type="checkbox"/>	772	583.63	ORH:	I've seen a	man	sell one for fourteen thousand dollar	586.08
<input type="checkbox"/>	1024	774.83	ORH:	No. One- another	man	helped me.	776.22
<input type="checkbox"/>	1418	1062.85	CM:	I would,	man	.	1063.77
<input type="checkbox"/>	1692	1274.99	ORH:	[I shot a]	man	.	1275.70
<input type="checkbox"/>	1890	1444.67	ORH:	one	man	that uh	1445.97
<input type="checkbox"/>	3238	2377.89	JF:	[Oh	man	]	2378.29

[ analyze selected lines from this transcript ] Check the boxes and click the link to the left to analyze select lines from this transcript

Transcript: **bee0030b\_0\_454** Click a line number to go to the transcript at that line [ 1 Result ]

Sel	Line#	Start T	Spkr		man		End T
<input type="checkbox"/>	328	252.88	MH:	straight- There comes the	man	you ought be [????]	256.53

[ analyze selected lines from this transcript ] Check the boxes and click the link to the left to analyze select lines from this transcript

Transcript: **bee0130a\_36\_218** Click a line number to go to the transcript at that line [ 1 Result ]

Sel	Line#	Start T	Spkr		man		End T
<input type="checkbox"/>	124	93.25	DY:	Old	man		93.71

[ analyze selected lines from this transcript ] Check the boxes and click the link to the left to analyze select lines from this transcript

Transcript: **dca\_alayna\_a\_0\_4245** Click a line number to go to the transcript at that line [ 11 Results ]

Sel	Line#	Start T	Spkr		man		End T
<input type="checkbox"/>	1644	1433.36	Alayna:	woman or that	man	strong as they grow up so they know more about their life, they know more,	1438.18

## Can create SLAAP-like custom interfaces

Transcript Summarizer - Beta

u.edu/ncslaap/transcript\_summary.php?t=w\_dca\_elisa\_a\_0\_2698

mail PHPman MySQLman FLAReNet EXMARaLDA Local & Kelly 1986 work-in-prog CorpLing LL-jobs

[ User Forum ] [ tskendal : Acct | Logout ] (autologout if no activity at 8:17:12)

marizer - Beta [ @transcript\_summary.php ] [ Library ]

Please delete some files or enable automatic cleanup.

atic Summarization - **dca\_elisa\_a\_0\_2698** What's this about? See Tyler's Working Paper (PDF, 1.3mb). [ View dca\_elisa\_a\_0\_2698 | All Transcripts ]

**Transcript Summary Statistics**

Transcript has **2** speakers: *Carissa, Elisa*

Transcript total temporal length: **2,697.89** seconds (44.96 minutes)

Transcript total line length: **2,916** lines (including blank lines, e.g., pauses)

Total non-blank lines: **1,458**

Speaker	Talk Lines <sup>1</sup>	Turn Lines <sup>1</sup>	Words	Words of Tran	Talk-Time (sec)	Talk-Time of Total Talk <sup>2</sup>	Turn-Time (sec)	Turn-Time of Entire Tran <sup>3</sup>
Carissa	626	980	3,172	41.76 %	728.06	42.13 %	966.44	35.82 %
Elisa	832	1,380	4,424	58.24 %	1,000.09	57.87 %	1,350.50	50.06 %
Totals:	1,458	2,360	7,596	100 %	1,728.15	100 %	2,316.94	85.88 %

<sup>1</sup> Talk Lines only include transcript lines with orthographic text. Turn Lines are all transcript lines that occur within a speaker's turn. The crucial difference between Talk Lines and Turn Lines is whether or not blank lines, or pauses, are counted. Blank lines are determined to "belong" to the speaker by occurring between two lines of talk. Talk-Time and Turn-Time are sums of the timespans of these two measurements of line "ownership".

<sup>2</sup> Talk-Time of Total Talk is the percentage of total talk (not including pauses) by each speaker. The sum of all the speakers' Talk-Time should always account for 100% of the total talk in the transcript.

<sup>3</sup> Turn-Time of Entire Tran is the percentage of how much of the entire duration of the transcript's time each speakers' total Turn-Time accounts for. The sum of this measure will usually be less than 100% as not all lines (namely, inter-turn pauses) "belong" to specific speakers. A high amount of speaker overlap (more overlap than inter-turn pause) can result in a result over 100%.

E.g.,

But could also access directly from remote MySQL clients

```
Terminal - mysql - 100x35
mysql> select * from trans_sbc047 where ortho regexp 'man';
+-----+-----+-----+-----+-----+
| line | speaker | st_t | ortho | en_t |
+-----+-----+-----+-----+-----+
| 2 | FRED | 1.500 | I tell you man, | 2.130 |
| 9 | FRED | 8.620 | .. that was <VOX just a little bit VOX> too long man. | 10.410 |
| 34 | FRED | 31.010 | oh= man, | 31.660 |
| 49 | RICHARD | 44.140 | [How many] cases you packed. | 45.520 |
| 49 | FRED | 45.520 | (H) I don't know man. | 46.550 |
| 52 | FRED | 48.600 | ... I don't know how many .. cases [that is], | 50.540 |
| 57 | FRED | 52.120 | .. that shit was heavy man. | 53.540 |
| 100 | FRED | 84.920 | I go look man, | 85.700 |
| 111 | FRED | 94.940 | aw man, | 95.560 |
| 112 | FRED | 95.560 | (H)= this is the pits man. | 97.240 |
| 126 | FRED | 107.420 | (H) It's p2]ar for the course man. | 109.000 |
| 150 | FRED | 128.440 | Th- the competition man. | 129.850 |
| 183 | RICHARD | 158.370 | I'll be up there too in the top four salesman. | 160.390 |
| 293 | FRED | 259.980 | ... (H) She still considers you man. | 262.240 |
| 370 | FRED | 339.340 | That's hard [man]. | 340.100 |
| 489 | RICHARD | 444.260 | w- I had a beautiful woman, | 445.330 |
| 635 | RICHARD | 594.840 | Let me go talk to my manager. | 596.050 |
| 739 | FRED | 705.210 | [I'm hip3] man, | 705.770 |
| 740 | FRED | 707.210 | @Man @@@, | 708.000 |
| 828 | FRED | 792.960 | So what does the Porsche have man. | 794.300 |
| 876 | FRED | 847.570 | Man, | 847.850 |
| 928 | FRED | 901.000 | ... That's gonna be a good show man, | 902.940 |
| 998 | FRED | 1026.910 | [That's cheap] man cause, | 1028.090 |
| 1005 | FRED | 1032.250 | .. % eleven-hundred too man. | 1033.760 |
| 1013 | FRED | 1035.900 | .. Aries man. | 1036.790 |
| 1114 | RICHARD | 1145.110 | ... I had a good woman, | 1147.900 |
+-----+-----+-----+-----+-----+
26 rows in set (0.00 sec)

mysql> _
```

```
mysql_concording.R
<functions> Help search

1 # Load R's MySQL support library
2 library(RMySQL)
3
4 # Connect to the database
5 mysql.con <- dbConnect(MySQL(), user="ruser", password="c0rpus",
6 dbname="exploring", host='localhost')
7
8 # E.g., Searching for words ending in "man"
9 search.term <- 'man'
10
11 # [[:~:] is MySQL for word end-boundary
12 regexp.term <- paste(search.term, '[[:~:]', sep="")
13
14 # Retrieve all matches from the database
15 db.matches <- dbGetQuery(mysql.con, paste("select * from trans_sbc047 where
16 ortho regexp '", regexp.term, "'"))
17
18 # Very simple output - convert all text to lowercase, match to upper
19 out1 <- gsub(paste("(\\w*", tolower(search.term), "\\b)", sep=""), "\\U\\1",
20 tolower(db.matches$ortho), perl=TRUE)
21
22 # Display results
23 out1
24
25 # Disconnect from the database
26 dbDisconnect(mysql.con)
```

```
R Console
> # Display results
> out1
[1] "i tell you MAN,"
[2] ".. that was <vox just a little bit vox> too long MAN."
[3] "oh= MAN,"
[4] "(h) i don't know MAN."
[5] ".. that shit was heavy MAN."
[6] "i go look MAN,"
[7] "aw MAN,"
[8] "(h)= this is the pits MAN."
[9] "(h) it's p2]ar for the course MAN."
[10] "th- the competition MAN."
[11] "i'll be up there too in the top four SALESMAN."
[12] "... (h) she still considers you MAN."
[13] "that's hard [MAN]."
[14] "w- i had a beautiful WOMAN,"
[15] "[3i'm hip3] MAN,"
[16] "@MAN @@@,"
[17] "so what does the porsche have MAN."
[18] "MAN, "
[19] "... that's gonna be a good show MAN,"
[20] "[that's cheap] MAN cause,"
[21] ".. % eleven-hundred too MAN."
[22] ".. aries MAN."
[23] "... i had a good WOMAN,"
>
> # Disconnect from the database
```

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.