Mobilizing Smaller Datasets for Large-Scale Phonetic Analysis: Web-Databases and Semi-Automatic Analyses

Tyler Kendall[1], *Ann R. Bradlow*[2]

[1] Department of Linguistics, University of Oregon, U.S.A.
[2] Department of Linguistics, Northwestern University, U.S.A.
tsk@uoregon.edu, abradlow@northwestern.edu

Abstract

In this paper, we consider what can be gained by aggregating smaller, individual recording collections into large, user-centric web-based archives. We review two web-based archiving projects, the Online Speech/Corpora Archive and Analysis Resource (OSCAAR) and the Sociolinguistic Archive and Analysis Project (SLAAP), both of which feature organization and analysis tools that connect and enhance the overall usefulness of the archived recordings. We also briefly provide an example from research on speech timing that illustrates how these tools can help aggregate across research collections and empower new analyses.

Index Terms: speech databases, audio archives, speech timing

1. Introduction

Recordings of speech comprise the backbone of acoustic phonetic, speech perception, and much general linguistic research. The creation of these speech recordings (e.g., recording equipment and its use) and their analysis (e.g., appropriate statistical methods) tend to be thoroughly treated in methodology sections of papers and in specialized methodology textbooks. However, the storage, management, and preservation of these resources are rarely discussed in the academic literature, though these practices influence both the short term and long term usability of these resources. Improving our data storage and management strategies allow us to leverage new technologies so that our data archives become not just usable, but maximally useful.

In many areas of linguistic and speech science research a large amount of time, energy, and cost goes into the recording and development of specialized datasets for the pursuit of specific research questions. For instance, sociolinguists often generate individual collections of field-recorded sociolinguistic interviews in the pursuit of their specific research questions, and phoneticians often engage in the design and collection of highly controlled, lab-based recordings for the same reasons. In both cases, fairly small numbers of subjects are traditionally recorded. As theories of speech and language production and perception have moved towards understanding the sources and consequences of intra- and inter-talker variability, and as computational and storage media resources have evolved, greater emphasis has been placed on the compilation of larger digitized speech databases. Such multi-talker databases support increasingly detailed and sophisticated sociophonetic and acoustic phonetic analyses and provide stimuli for speech perception experiments that directly address the nature and extent of speech signal variability. Nevertheless, the typical digitized speech recording dataset collected within the context of linguistic research still cannot be considered very large scale and there is very limited aggregation across isolated datasets. Many papers have recently argued (e.g., [1], [2]) for improving methods of linguistic data archiving and preservation so that diverse recording collections can be aggregated and mobilized for the pursuit of new, broader, deeper and often unplanned for research questions.

In this paper, we present two related web-based archiving projects, the Sociolinguistic Archive and Analysis Project (SLAAP) and the Online Speech/Corpora Archive and Analysis Resource (OSCAAR). Each of these web-based resources features organization and analysis tools that enhance the overall usefulness of the archived recordings. Crucially, they allow researchers to span across collections of recordings and annotations from different projects in order to mobilize disparate, smaller datasets for large-scale phonetic research. To illustrate this, and to exemplify some substantive outcomes of these projects, we conclude the paper by briefly presenting some new approaches to investigating sociolinguistic and sociophonetic variation from a large-scale research project, which has been made possible by SLAAP's software.

2. Web-based speech archives

While the past several decades have seen a massive increase in the availability of public collections of language data for research purposes (thanks in no small part to groups like the Linguistic Data Consortium [http://ldc.upenn.edu/] and the TalkBank project [http://talkbank.org/]), most emphasis in the publication and dissemination of language data resources has been placed on the development of *individual* corpora. The TalkBank.org website is one major exception in that it presents an umbrella interface over a large range of individual data collections [3]. OSCAAR and SLAAP, two projects that we have been involved in, we believe, represent another set of exceptions, and we turn to describe these now. We begin with OSCAAR, the newer of the two projects. (SLAAP is given a briefer treatment here as it has been described in several publications, e.g., [2], [4], [5].)

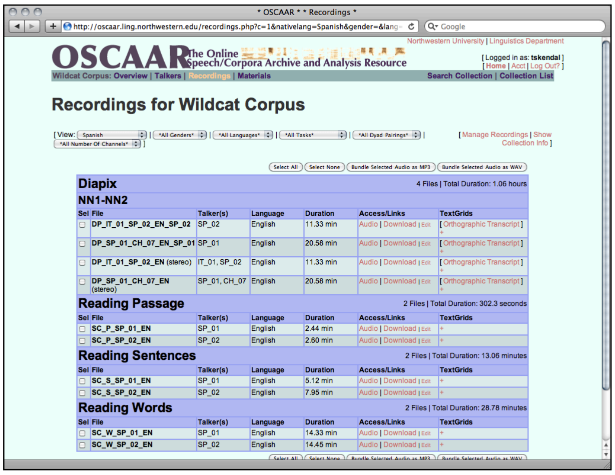
2.1. OSCAAR

OSCAAR, the Online Speech/Corpora Archive and Analysis Resource [http://oscaar.ling.northwestern.edu/], was begun in the Fall of 2009 in the Speech Communication Research Lab of the Northwestern University Linguistics Department [6]. OSCAAR seeks to provide a web-based storage and access facility for speech recordings from numerous talkers, languages, speaking styles, and elicitation methods that have accumulated over the past decade of research in the lab. OSCAAR allows for the flexible organization of various kinds of speech recordings so that researchers can more easily

- retrieve and review possible speech recordings for acoustic or other linguistic analysis (as in Figure 1);
- retrieve and review possible speech recordings for use as stimuli in speech recognition and perception experiments;

- search across recording collections to access, compare, analyze, etc. recordings from multiple collections
- gain new perspectives on their data (see Figures 4 & 5 for examples); and, of course,
- preserve recording collections, and share them, via the web-interface, with colleagues (while following security and access procedures that respect talker privacy as specified in IRB-approved consent forms).

While still in its early phases of development, OSCAAR already houses many of the speech recording collections developed by the Speech Communication Research Lab and we are actively adding data collections from other researchers. At the time of this writing, there are approximately 45,000 audio recordings stored in OSCAAR. These range from one-second long recordings of individual words and sentences to 30-minute long recordings of task-based spontaneous speech. As an example of OSCAAR's interface, Figure 1 displays a screenshot of English language recordings in the Wildcat Corpus [7] from native Spanish language talkers. Users can view, explore, and search the recordings in a variety of other ways, from characteristics of the talker, to collection-specific organizational terms (the "Diapix", "NN1-NN2", "Reading Passage", "Reading Sentences", and "Reading Words" headers in Figure 1 are all custom, sort-able filter-able organizational terms within the Wildcat Corpus; other collections can use other terms and categories), and, finally, to complex searches across multiple collections.


* OSCAAR * * Recordings *
http://oscaar.ling.northwestern.edu/recordings.php?c=1&nativelang=Spanish&gender=&lang=
Google
Northwestern University | Linguistics Department
OSCAAR The Online Speech/Corpora Archive and Analysis Resource
[Logged in as: tskendal]
[Home | Acct | Log Out?]
Wildcat Corpus: Overview | Talkers | Recordings | Materials
Search Collection | Collection List
[Manage Recordings | Show Collection Info]
Select All
Select None
Bundle Selected Audio as MP3
Bundle Selected Audio as WAV

Recordings for Wildcat Corpus

[View: Spanish | *All Genders* | *All Languages* | *All Tasks* | *All Dyad Pairings* | *All Number Of Channels*]

Diapix — 4 Files | Total Duration: 1.06 hours

NN1-NN2

Sel	File	Talker(s)	Language	Duration	Access/Links	TextGrids
□	DP_IT_01_SP_02_EN_SP_02	SP_02	English	11.33 min	Audio | Download | Edit	[Orthographic Transcript] +
□	DP_SP_01_CH_07_EN_SP_01	SP_01	English	20.58 min	Audio | Download | Edit	[Orthographic Transcript] +
□	DP_IT_01_SP_02_EN (stereo)	IT_01, SP_02	English	11.33 min	Audio | Download | Edit	[Orthographic Transcript] +
□	DP_SP_01_CH_07_EN (stereo)	SP_01, CH_07	English	20.58 min	Audio | Download | Edit	[Orthographic Transcript] +

Reading Passage — 2 Files | Total Duration: 302.3 seconds

Sel	File	Talker(s)	Language	Duration	Access/Links	TextGrids
□	SC_P_SP_01_EN	SP_01	English	2.44 min	Audio | Download | Edit	+
□	SC_P_SP_02_EN	SP_02	English	2.60 min	Audio | Download | Edit	+

Reading Sentences — 2 Files | Total Duration: 13.06 minutes

Sel	File	Talker(s)	Language	Duration	Access/Links	TextGrids
□	SC_S_SP_01_EN	SP_01	English	5.12 min	Audio | Download | Edit	+
□	SC_S_SP_02_EN	SP_02	English	7.95 min	Audio | Download | Edit	+

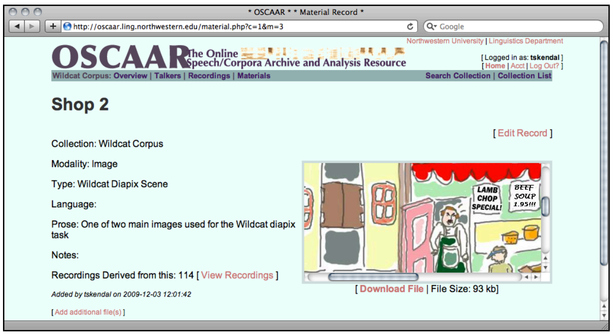
Reading Words — 2 Files | Total Duration: 28.78 minutes

Sel	File	Talker(s)	Language	Duration	Access/Links	TextGrids
□	SC_W_SP_01_EN	SP_01	English	14.33 min	Audio | Download | Edit	+
□	SC_W_SP_02_EN	SP_02	English	14.45 min	Audio | Download | Edit	+

Figure 1: *Screenshot of OSCAAR recordings view*

In addition to the speech recordings themselves, OSCAAR also stores information about the talkers, and the associated materials for those recordings, such as the reading passages, interview protocols, or other stimuli used in eliciting the speech. These can be reviewed within OSCAAR and users can quickly peruse the recordings derived from a given material. This is demonstrated in Figure 2, a picture used for the "Diapix" task in the Wildcat Corpus [7].

Figure 3 displays an inline audio player for one of the Wildcat Corpus recordings derived from the picture shown in Figure 2. Here a user can create – and later search and/or return to – time-stamped notes that can be associated to specific moments in an audio recording. Directly from the audio player page, users can also review the stimuli used to elicit the recording (if any) and "zoom in" and view a spectrogram for a short span of the audio. Users can also download complete audio files or extract portions of the audio by entering a time-range.


* OSCAAR * * Material Record *
http://oscaar.ling.northwestern.edu/material.php?c=1&m=3
Google
Northwestern University | Linguistics Department
OSCAAR The Online Speech/Corpora Archive and Analysis Resource
[Logged in as: tskendal]
[Home | Acct | Log Out?]
Wildcat Corpus: Overview | Talkers | Recordings | Materials
Search Collection | Collection List
[Edit Record]
LAMB CHOP SPECIAL!
BEEF SOUP 1.95!!!
[Download File | File Size: 93 kb]
[Add additional file(s)]

Shop 2

Collection: Wildcat Corpus

Modality: Image

Type: Wildcat Diapix Scene

Language:

Prose: One of two main images used for the Wildcat diapix task

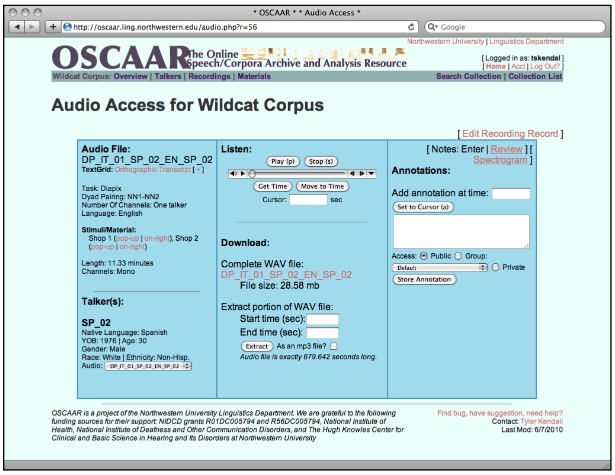
Notes:

Recordings Derived from this: 114 [View Recordings]

Added by tskendal on 2009-12-03 12:01:42

Figure 2: *OSCAAR interface to a picture stimuli for the Wildcat Corpus*

Additionally, OSCAAR provides a static URL and a "homepage" (not shown) for each data collection, so that researchers can publish persistent links to their dataset and then chose how much access to give public viewers. Visitors to OSCAAR can view basic information about data collections, but must be granted access by the site's administrator and the collection's owner to access each individual recording collection.


* OSCAAR * * Audio Access *
http://oscaar.ling.northwestern.edu/audio.php?r=56
Google
Northwestern University | Linguistics Department
OSCAAR The Online Speech/Corpora Archive and Analysis Resource
[Logged in as: tskendal]
[Home | Acct | Log Out?]
Wildcat Corpus: Overview | Talkers | Recordings | Materials
Search Collection | Collection List
[Edit Recording Record]
Play (p)
Stop (s)
Get Time
Move to Time
Set to Cursor (a)
Store Annotation
Extract
Find bug, have suggestion, need help?

Audio Access for Wildcat Corpus

Audio File:
DP_IT_01_SP_02_EN_SP_02
TextGrid: Orthographic Transcript [+]

Task: Diapix
Dyad Pairing: NN1-NN2
Number Of Channels: One talker
Language: English

Stimuli/Material:
Shop 1 (pop-up | on-right), Shop 2 (pop-up | on-right)

Length: 11.33 minutes
Channels: Mono

Talker(s):

SP_02
Native Language: Spanish
YOB: 1976 | Age: 30
Gender: Male
Race: White | Ethnicity: Non-Hisp.
Audio: DP_IT_01_SP_02_EN_SP_02

Listen:

Cursor: sec

Download:

Complete WAV file:
DP_IT_01_SP_02_EN_SP_02
File size: 28.58 mb

Extract portion of WAV file:
Start time (sec):
End time (sec):
As an mp3 file? □
Audio file is exactly 679.642 seconds long.

[Notes: Enter | Review] [Spectrogram]

Annotations:

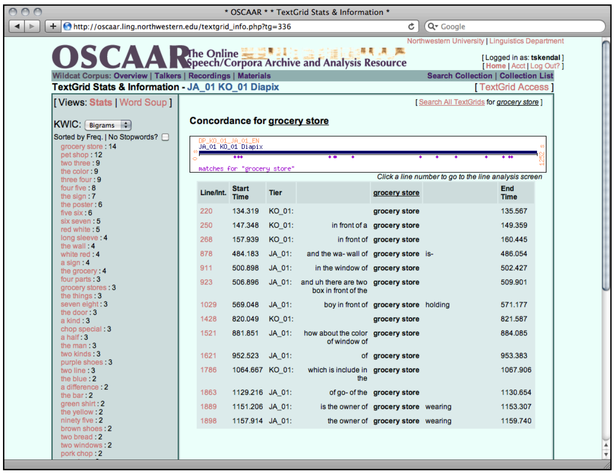
Add annotation at time:

Access: ● Public ○ Group: Default ○ Private

OSCAAR is a project of the Northwestern University Linguistics Department. We are grateful to the following funding sources for their support: NIDCD grants R01DC005794 and R56DC005794, National Institute of Health, National Institute of Deafness and Other Communication Disorders, and The Hugh Knowles Center for Clinical and Basic Science in Hearing and Its Disorders at Northwestern University

Contact: Tyler Kendall
Last Mod: 6/7/2010

Figure 3: *Audio access within OSCAAR*


* OSCAAR * * TextGrid Stats & Information *
http://oscaar.ling.northwestern.edu/textgrid_info.php?tg=336
Google
Northwestern University | Linguistics Department
OSCAAR The Online Speech/Corpora Archive and Analysis Resource
[Logged in as: tskendal]
[Home | Acct | Log Out?]
Wildcat Corpus: Overview | Talkers | Recordings | Materials
Search Collection | Collection List
[TextGrid Access]
[Views: Stats | Word Soup]
[Search All TextGrids for grocery store]
DP_KO_01_JA_01_EN
JA_01 KO_01 Diapix
0 s
1252 s
matches for "grocery store"

TextGrid Stats & Information - JA_01 KO_01 Diapix

KWIC: Bigrams
Sorted by Freq. | No Stopwords? □

grocery store	14
pet shop	12
two three	9
the color	9
three four	9
four five	8
the sign	7
the poster	6
five six	6
six seven	5
red white	5
long sleeve	4
the wall	4
white red	4
a sign	4
the grocery	4
four parts	3
grocery stores	3
the things	3
seven eight	3
the door	3
a kind	3
chop special	3
a half	3
the man	3
two kinds	3
purple shoes	3
two line	3
the blue	2
a difference	2
the bar	2
green shirt	2
the yellow	2
ninety five	2
brown shoes	2
two bread	2
two windows	2
pork chop	2

Concordance for grocery store

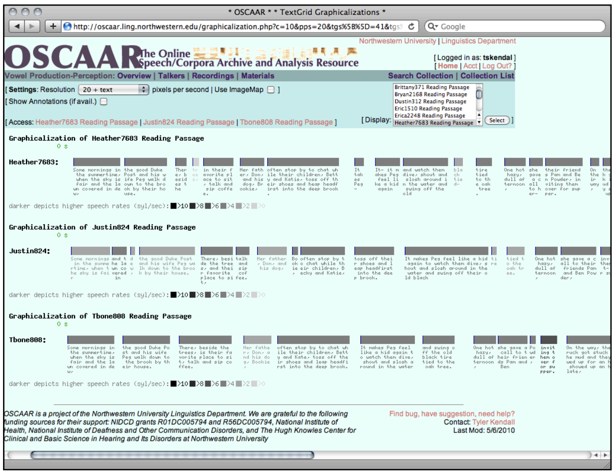
Click a line number to go to the line analysis screen

Line/Int.	Start Time	Tier		grocery store		End Time
220	134.319	KO_01:		grocery store		135.567
250	147.348	KO_01:	in front of a	grocery store		149.359
268	157.939	KO_01:	in front of	grocery store		160.445
878	484.183	JA_01:	and the wa- wall of	grocery store	is-	486.054
911	500.898	JA_01:	in the window of	grocery store		502.427
923	506.896	JA_01:	and uh there are two box in front of the	grocery store		509.901
1029	569.048	JA_01:	boy in front of	grocery store	holding	571.177
1428	820.049	KO_01:		grocery store		821.587
1521	881.851	JA_01:	how about the color of window of	grocery store		884.085
1621	952.523	JA_01:	of	grocery store		953.383
1786	1064.667	KO_01:	which is include in the	grocery store		1067.906
1863	1129.216	JA_01:	of go- of the	grocery store		1130.654
1889	1151.206	JA_01:	is the owner of	grocery store	wearing	1153.307
1898	1157.914	JA_01:	the owner of	grocery store	wearing	1159.740

Figure 4: *Concordance view with time information*

For transcribed data in OSCAAR, users have access to corpus-like features that are enhanced through the time-alignment built into the implementation of the transcripts (see §3). Figure 4, for instance, displays a concordance for "grocery store" for one of the Wildcat Corpus diapix task recordings. Figure 4 also automatically shows us that "grocery store" is the most frequent two-word collocation in the recording, and displays, in addition to a normal concordance view of the text, the time-stamps of each relevant utterance and their location on a graphical time-line of the recording.

OSCAAR further provides a range of advanced features that provide new insight into spoken language recording collections. For instance, Figure 5, using data from [8], provides one example of a way that OSCAAR dynamically provides new ways to interface with and view speech data.


* OSCAAR * * TextGrid Graphicalizations *
http://oscaar.ling.northwestern.edu/graphicalization.php?c=10&pps=20&tgs%5B%5D=41&tgs
Google
Northwestern University | Linguistics Department
OSCAAR The Online Speech/Corpora Archive and Analysis Resource
[Logged in as: tskendal]
[Home | Acct | Log Out?]
Vowel Production-Perception: Overview | Talkers | Recordings | Materials
Search Collection | Collection List
Brittany371 Reading Passage
Bryan2168 Reading Passage
Dustin312 Reading Passage
Eric1510 Reading Passage
Erica2248 Reading Passage
Heather7683 Reading Passage
[Display:
Select]
Graphicalization of Heather7683 Reading Passage
0 s
Heather7683:
Graphicalization of Justin824 Reading Passage
0 s
Justin824:
Graphicalization of Tbone808 Reading Passage
0 s
Tbone808:
Find bug, have suggestion, need help?
Contact: Tyler Kendall
Last Mod: 5/6/2010

[Settings: Resolution 20 + text pixels per second | Use ImageMap □]

[Show Annotations (if avail.) □

[Access: Heather7683 Reading Passage | Justin824 Reading Passage | Tbone808 Reading Passage]

darker depicts higher speech rates (syl/sec): ■>10 ■>8 ■>6 ■>4 ■>2 ■>0

OSCAAR is a project of the Northwestern University Linguistics Department. We are grateful to the following funding sources for their support: NIDCD grants R01DC005794 and R56DC005794, National Institute of Health, National Institute of Deafness and Other Communication Disorders, and The Hugh Knowles Center for Clinical and Basic Science in Hearing and Its Disorders at Northwestern University

Figure 5: *OSCAAR graphicalization of aligned orthographic text for two talkers reading the same reading passage*

This *graphicalization* (cf. [4]) depicts the speech timing of talkers reading (aloud) the same reading passage. The darkness of shading of each gray block indicates the rate of speech (automatically determined by the software) and the width of each block its duration. Blank areas reflect silent pauses, with the width of the blank area depicting the duration of the silent pause. The version displayed in Figure 5 is set to a resolution of 20 pixels per second and to display the utterances' time-aligned textual representations beneath each block. These kinds of displays are widely customizable by OSCAAR's users.

2.2. The Sociolinguistic Archive and Analysis Project

While OSCAAR is designed to be a general-purpose speech archive, SLAAP, an older project, was developed for specifically sociolinguistic purposes. SLAAP is housed at North Carolina State University and is jointly supported by the NCSU Libraries and the North Carolina Language and Life Project, a research and outreach project led by Walt Wolfram and colleagues at NCSU. SLAAP features a growing archive of sociolinguistic audio recordings along with dynamic interfaces to those recordings. Many of the features are similar to those described above for OSCAAR, though SLAAP, since it has existed longer and has been designed around a specific research focus has additional tools developed for sociolinguistic analysis.

At the time of this writing, over 1,600 interview recordings are stored in and accessible through SLAAP, amounting to over 1,300 hours of recorded speech. The centerpiece of the SLAAP software is a time-aligned annotation framework that is integrated with analytic software (including Praat [9] and R [10]) allowing for features like the automatic generation of spectrograms within the web-based audio player, the extraction of phonetic data from within a recording's transcript, multiple and dynamic displays of each transcript, and corpus linguistic analyses across the diverse materials in the archive. At present (only) 3% of the total archive is transcribed. This constitutes a 385,000 word searchable corpus, representing 38.5 hours of recorded talk, time-aligned at the utterance level. Like the digitization and entry of the audio recordings, transcription is ongoing, though – as is evident from the small percentage of transcribed talk – is more slow-going than digitization. (SLAAP and its features are discussed more fully in [4], [11], and [12].)

3. Data-based and Praat-based annotations

Both OSCAAR and SLAAP have been designed around a data-based time-aligned annotation model where all annotation layers reside in a MySQL relational database. While the systems themselves are agnostic about how annotation types like transcription and aligned text are developed, the most common formats for these to originate in are Praat TextGrids [9]. Praat TextGrids allow flexible kinds of annotation for audio recordings with arbitrarily fine temporal precision. This allows both gross and extremely fine time-alignment, and types of annotation ranging from phonetic and phone-level transcription to grammatical mark up and coding to orthographic and utterance-level broad transcription.

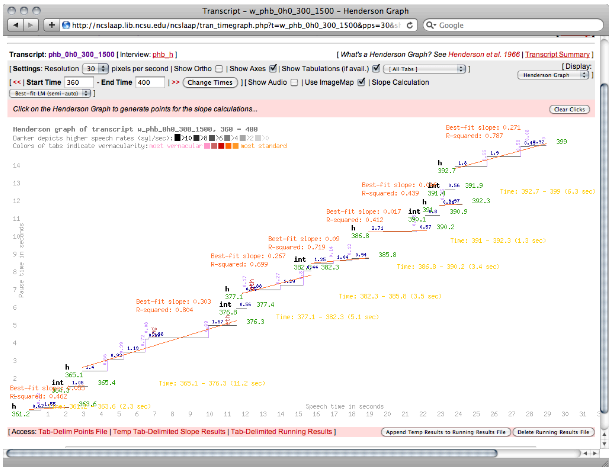
Once created, TextGrids can be uploaded and attached to their source audio in both OSCAAR and SLAAP. The systems parse the TextGrid files and store the annotations as relational database data. Users can export the annotations out of the system back to a TextGrid format or as tab-delimited text. While not a direct implementation of the *annotation graph* framework [13], the model developed here bears many similarities.

Relational databases provide an extremely fast and stable backend for web-based corpora. Davies [14], for example, has demonstrated that relational databases make efficient storage and search engines even for extremely large corpora.

4. Aggregating across collections

One major advantage, we argue, to the web- and data-based approach we have taken to managing speech data collections is that software tools, like features in OSCAAR and SLAAP or other, customized analysis scripts, can reach across independent recording collections to perform analyses spanning across the boundaries of the projects from which the original datasets came.

For instance, in [11], SLAAP was used to conduct a corpus sociophonetic analysis of speech timing across fourteen separate sociolinguistic research projects. In addition to examining variation in articulation rate and silent pause durations, which were measured by mining the databased, time-aligned transcripts, a software-based plotting method was also developed that allowed for the investigation of the influence of *hesitancy* on the realization of morphosyntactic and phonological sociolinguistic variables. This method involves the generation of *Henderson graphs* (cf. [11] and [15]) and the use of the Henderson graphs' *slopes* as an independent predictor in the statistical analysis of other variables. Figure 6 displays a screenshot of a Henderson graph in SLAAP (data come from [16]).



Transcript – w_phb_0h0_300_1500 – Henderson Graph

http://ncslaap.lib.ncsu.edu/ncslaap/tran_timegraph.php?t=w_phb_0h0_300_1500&pps=30&s

Transcript: phb_0h0_300_1500 [Interview: phb_h]

[What's a Henderson Graph? See Henderson et al. 1966 | Transcript Summary]

[Settings: Resolution 30 pixels per second | Show Ortho □ | Show Axes ☑ | Show Tabulations (if avail.) ☑ [All Tabs]]

[<< | Start Time 360 - End Time 400 | >> Change Times] [Show Audio □ | Use ImageMap ☑ | Slope Calculation Best-fit LM (semi-auto)]

[Display: Henderson Graph]

Click on the Henderson Graph to generate points for the slope calculations...

Henderson graph of transcript w_phb_0h0_300_1500, 360 – 400

Darker depicts higher speech rates (syl/sec): >10 >8 >6 >4 >2 >0

Colors of tabs indicate vernacularity: most vernacular most standard

[Access: Tab-Delim Points File | Temp Tab-Delimited Slope Results | Tab-Delimited Running Results]

Figure 6: *SLAAP Henderson graph showing best-fit slope lines and the occurrence of sociolinguistic variables*

The Henderson graph slopes prove to be a useful independent variable in statistical models of several sociolinguistic variables and allow for the testing of hypotheses, new and old [17], about the relationship between paralinguistic cues, like hesitation phenomena, and the realization of variable forms.

For instance, in examining the productions of variable (ing), the alternation between [ɪn] and [ɪŋ] in word-final *–ing* in English, the slope of the stretch of talk containing each (ing) is found to be a statistically significant predictor in logistic regressions. Steeper slopes – representing a quantitative and systematic measure of more hesitant speech – are more likely to be realized as full [ɪŋ] than shallower slopes. Importantly, for the purposes of this paper, we observe that through SLAAP's software and collective archive, we can readily test (and have tested) this analysis against multiple recording collections.

5. Conclusion

In closing, we argue that web-based archives such as OSCAAR and SLAAP, which seek to develop organizational systems for multiple recording collections allow for the collected recordings to be better preserved, more accessible, and better mobilized for larger, aggregated analysis. These resources are particularly important for leveraging the full analytic potential of lab-based and interview-based speech and language corpora, however, this basic design approach also holds significant potential for very large-scale phonetic research based on corpora captured from broadcast and other public media such as radio, television and film. We believe that storage, management and preservation tools like these are an essential complement to the very-large-scale corpus analysis tools and methodologies that are increasingly available to speech and language researchers and innovators.

6. Acknowledgements

OSCAAR has been developed thanks to funding from the NIH, grant NIDCD R01 DC005794 (A. R. Bradlow, PI). SLAAP has been supported by the NCSU Libraries and the William C. Friday Endowment at NCSU. We are grateful to Kelsey Mok, Chun Chan, Walt Wolfram, Erik Thomas, and many more individuals than we can list here for help with the development of OSCAAR and SLAAP and the creation of their many archived recording collections.

7. References

[1] Bird, S. and Simons, G., "Seven Dimensions of Portability for Language Documentation and Description", *Language* 79(3), 557-82, 2003.

[2] Kendall, T., "On the History and Future of Sociolinguistic Data", *Language and Linguistics Compass* 2(2), 332-351, 2008.

[3] MacWhinney, B., "The TalkBank Project", in J. Beal, K. Corrigan, and H. Moisl, Creating and Digitizing Language Corpora, Volume 1, 163-180, Palgrave-Macmillan, 2007.

[4] Kendall, T., "The North Carolina Sociolinguistic Archive and Analysis Project: Empowering the Sociolinguistic Archive", *Penn Working Papers in Linguistics* 13(2), 15-26, 2007.

[5] Newman, J., "Spoken Corpora: Rationale and Application", *Taiwan Journal of Linguistics* 6(2), 27-58, 2008.

[6] Kendall, T., "Developing Web Interfaces to Spoken Language Data Collections", Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science, 2010.

[7] Van Engen, K., Baese-Berk, M., Baker, R., Choi, A., Kim, M., and Bradlow, A. R., "The Wildcat Corpus of Native- and Foreign-Accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles", *Language and Speech*, forthcoming.

[8] Kendall, T. and Fridland, F., "Mapping Production and Perception in Regional Vowel Shifts", *Penn Working Papers in Linguistics* 16(2), art. 13, 2010.

[9] Boersma, P. and Weenink, D., "Praat: Doing Phonetics by Computer", 2010. [Computer Program]

[10] R Development Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria, 2010. [Computer Language]

[11] Kendall, T., "Speech Rate, Pause, and Linguistic Variation: An Examination Through the Sociolinguistic Archive and Analysis Project", Doctoral Dissertation, Duke University, 2009.

[12] Kendall, T., "SLAAP User Guide, Version 0.96", Online: http://ncslaap.lib.ncsu.edu/userguide/

[13] Bird, S., and Liberman, M. "A Formal Framework for Linguistic Annotation", *Speech Communication* 33(1-2), 23-60, 2001.

[14] Davies, M., "The Advantage of Using Relational Databases for Large Corpora: Speed, Advanced Queries, and Unlimited Annotation", *International Journal of Corpus Linguistics* 10, 301-328, 2005.

[15] Henderson, A., Goldman-Eisler, F., and Skarbek, A., "Sequential Temporal Patterns in Spontaneous Speech", *Language and Speech* 9(4), 207-216, 1966.

[16] Childs, R., De Decker, P., Deal, R., Kendall, T., Thorburn, J., Williamson, M., and Van Herk, G., "Stop Signs: The Intersection of Interdental Fricatives and Identity in Newfoundland", *Penn Working Papers in Linguistics* 16(2), art. 5, 2010.

[17] Labov, W. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.