DIALECT EVOLUTION AND ONGOING VARIABLE LINGUISTIC INPUT: PRODUCTION AND PERCEPTION OF THE ENGLISH SPOKEN IN THE PACIFIC NORTHWEST

NSF Grant #: BCS-1147678, 0643374
UW ID (used by UW Grants and Contracts, Office of Sponsored Projects): A25753
UW IRB ID: 42736

Grant Funding Dates: June 15, 2012 – November 30, 2014

public website https://zeos.ling.washington.edu/~PNWEnglish/
collaboration space zeos.ling.washington.edu

PNWE Analysis Handbook

(revised: October 2018, ABW)

University of Washington, Department of Linguistics
National Science Foundation
The **Pacific Northwest English Study**

Contents

I. PROJECT OVERVIEW

A. Team roles:

Title	Name (email)	Role
Principal Investigator	Alicia Beckford Wassink (wassink@uw.edu)	-Grant activity oversight -Data collection -Data analysis
Graduate Research Assistants	John Riebold Lisa Tittle *Meghan Oxley* *Robert Sykes* *Rachel Schirra* *Michael Scanlon* *Manuela Rasing*	-Field data collection -Data analysis -Inter-annotator reliability checking -Perception study design assistance
Undergraduate Research Assistants	Macklin Blackburn Anna Kunz Huayi Jiang *Chang Liu* *Samantha M. Sanches* *Rose Cooper-Finger* *Anonymous Senior* *Robert Squizzero*	-Scheduling and recruitment -Metadata markup -Orthographic transcription -Fieldwork assistance -Perception test data analysis
Linguistics Department Administrator	Michael Furr (furr@uw.edu)	-Personnel Administration -Time Reporting -Budget Administration
Computing Specialist	David Nichols (linglab@uw.edu)	-Purchasing -System administration -Omnilock access

B. Calendar:

Yr 1	SUM	Locations: Seattle, Yakima Valley PI – Establish budgets & payroll accounts, complete hiring, provide orientation in study protocols for Graduate and Undergraduate researchers, collect data for 5-7 Af-Am speakers. UGV – Study advertisement, respondent recruitment, scheduling, assist with collection of data for 10-15 H-Am speakers REU – Scheduling support, metadata work, collect data for 5-7 Af-Am RA1,2 – Assist with stimulus modification, collect data for 10-15 N-Am speakers, mentoring of REU RA3 – Assist with stimulus modification, collect data for 10-15 H-Am speakers
	AUT, WIN	Locations: Seattle PI – Grant administration, field team supervision, mentoring & annual evaluations RA1 – Collect data for. 6-7 Af-Am, + As-Am speakers, mentor REU RA2 – Collect data for. 6-7 Af-Am + As-Am speakers REU – Scheduling, web support, extract PNWE I data for perception study UGR – Transcribe conversational recordings
	SPR	Location: Tri-Cities PI – Grant administration, field team supervision, mentoring, perception test design RA1,2 – Collect data for 10-15 Af-Am speakers REU – Scheduling, Metadata work, some acoustic analysis
Yr 2	SUM	Locations: Tri-Cities and Spokane PI – Develop perception study online interface, team supervision, mentoring RA1,2 – Collect production data for 5-7 H-Am + 5-7 Af-Am speakers RA3 – Collect production data for 10-15 N-Am speakers REU – Generate recruitment ads for perception study, monitor online data collection UGR – Transcribe Nat-Am & H-Am conversational recordings
	AUT, WIN	Remote locations: none PI—Grant administration, Team supervision, mentoring RA1—Monitor online perception study, Acoustic analysis RA2—Perception study data summarization REU—Acoustic analysis, web support UGR—Transcribe C-Am conversational recordings
	SPR	PI—Team supervision, mentoring, perception data aggregation and analysis RA1,2—Perception data analysis

Legend:
PI: Principal Investigator
RA1-3: Graduate Research Assistants
UGV: Undergraduate Research Volunteer (Summer, year 1 only)
UGR: Undergraduate Research Assistant (LING499)
REU: Research Experience for Undergraduates position

C. Equipment:

Type	Make/Model	Accessories
Digital Flash Recorders	Samson H4 Zoom	compact flash memory, earphones
Telephone recording devices	JK Audio Host-desktop digital hybrid	M-Audio computer software for interfacing with MAC OSX
Microphone (if necessary)	Audio-Technica 3031 Tabletop microphone Small-diaphragm Cardioid Condenser Mic with 30Hz-20kHz Frequency Range and 158dB SPL Capacity	Microphone stand
Telephone headset	Sennheiser HMD281XQ	1-ear headset with microphone
Analysis computer	Apple iMac	
I/O interface	M-Audio Firewire 610	
Elicitation	Apple iPad 1	earphones, bluetooth keyboard

D. Software needed for different analysis-related tasks and analysis overview:
Our analysis protocols are summarized in Section C of the National Science Foundation proposal document. The first workshop will only tackle steps 1-3 for now, and later workshops will handle other steps):

What makes this a sociophonetic study? The instrumental phonetic analysis that brings acoustic information to the diagnosis of dialect forms is still novel in sociolinguistics, and the attention to detailed demographic and social network information is not the norm in experimental phonetics. This makes the study's data rich, but also very complex. We have a range of software tools at our disposal that handle pieces of the analysis. Here is a summary of how we will divide the labor among different software tools:

Task	Software
Generate conversation transcripts	**ELAN**
Word/vowel-level tagging in transcripts	**ELAN**
Vowel analysis	**PNWE Arpabet Vowel Analyzer**
Export data for by-speaker summarization of acoustic measures (f0, F1, F2, F3, duration)	**PNWE Arpabet Vowel Analyzer**
Phone/word-level forced alignment	**P2FA**
By-speaker summarization of acoustic data	**Excel**
Normalization, statistical testing	**R, SPSS**
Generate by-speaker, by-group graphs	**phonR** **Akustyk (**basic vowel quadrilaterals) **VOIS3D** (vowel ellipses, overlap calculation)
Data summarization, indexation with speaker variables	**FileMaker Pro**
Network graphing	**R**

II. General Equipment & Software Use

A. Recording Equipment:

Fig. 1: H4n Zoom flash recorder with built-in microphones
(peripherals: batteries, a/c adapter, earbuds, stand)

(a) test function with mic-to-line amplifier
(b) test function with pre-amplification (no mic-to-line amplifier)

B. Presentation of Elicitation Materials:

Fig. 2: Verizon JetPack 4510L 4G Hotspot

Fig. 3: iPad 1 (peripherals: bluetooth keyboard; software: FileMaker Go app)

(a) to display online consent form
(b) display elicitation materials

C. Using an iPad to Access the FMP Database:
Note: To be sharable, Filemaker Pro database needs to be FMPro v.12 (or newer)

We want to be able to access the demographic database in the field. We will enter demographic information directly into the database, rather than using the pen and paper method. Changes made to the database using the mobile application, FileMaker-Go, will be automatically change the contents of the database on Chesterton. Here is how to connect to database using the iPad:

1. To prepare the database on Chesterton (or other host machine):
    a. Most recent version of the database must be downloaded from the ORCA.
    b. On transfer from the ORCA, sometimes the FileMaker database is confused for an Excel file. Just re-initiate the download and save it again.

2. Prepare the database file in FileMaker Pro 12.0:
    a. Select File > Sharing > Filemaker network
    b. Sharing window opens
    c. The desired database file should appear in list of open files
    d. Select "All users"
    e. turn Sharing to ON (no other changes should be needed for this screen)
    f. should be accessible for all users (but can pwd protect it)
    g. IP address for Chesterton: 128.95.69.205 (this is not a static url, and may change)
    h. access terminal window, and issue `ipconfig` at the labuser prompt to see ip address, if necessary

3. Access the database on an iPad:
    a. Install FileMaker Go (Grant has purchased licenses for this purpose)
    b. Launch FileMaker Go
    c. Should be able to select Chesterton in Favorite Hosts menu
    d. Touch database filename on screen in list below "Favorite Hosts"
    e. When prompted, enter login `admin` and password `0r3g0n#tr41l`

    Note: FileMaker Go allows you to change layouts, as well as may create a new record.

4. Creating a Record
    a. Click the +/- button and select "Add New Record"
    b. Select sample type
    c. Enter subject ID
    d. Enter the rest of the demographic information

D. Accessing the On-line Research Collaboration Area (ORCA):
The URL for the ORCA is:
http://www.orca.artsci.washington.edu/sites/NWEnglish/default.aspx

To gain access to this space you will need to have an account set up. To set up an account you must contact the PI of the grant.

In the sociolinguistics lab the desktop computers designated for PNWE research are Chesterton and Abbadon. There are PNWE logins on both of these computers.

The password for the PNWE login is: `pwnenglish`
The password for the undergrad research login is: `pnwe-student`

Any locally used files (e.g., associated with a current task) should be stored under the PNWE login, but as a general practice all grant related files should also be stored in the ORCA, since most accessing of files is done remotely. This means the ORCA is the main repository for all files. Desktops are just for working copies. Think of the ORCA as needing to remain up-to-date at all times.

E. File Workflow:
General workflows for dealing with grant related files should follow the following steps:

1. Checkout the file from ORCA
    a. Hover over file name
    b. A down arrow will appear to the right of the file name
    c. Click the down arrow and select Check Out from the dropdown menu

2. Download file from ORCA
    a. Option 1
        i. Hover over file name
        ii. A down arrow will appear to the right of the file name
        iii. Click down arrow and select Send To from the dropdown menu
        iv. Select Download a Copy from the submenu
    b. Option 2
        i. Click file icon – this should immediately initiate a download of the file.

3. Work on the file on your local hard drive

4. Upload and check in file to ORCA

5. Delete the working copy from your local hard drive

F. Accessing Sound Files:
The sound files are stored on a special file server: (zeos.ling.washington.edu)
(for backup data, we use zygon.ling.washington.edu)
username: `pnwe`
password: `sociolab#$`

To access Zygon you will need to use an SFTP client (e.g. Flow, Fugu, FileZilla, Fetch, WinSCP, or the terminal).

G. Setting up P2FA:
The Penn Phonetics Lab Forced Aligner (P2FA) is a Python script which takes a .wav file recording and a .txt file transcript and generates a time-aligned Praat TextGrid of all words and phones in the

data. Unfortunately however, P2FA is not terribly user-friendly or easy to setup. The following is a step-by-step guide to installing and running P2FA. (Note: these steps are for Windows users, but the procedure on Mac and Linux machines is very similar and I will note differences where possible). See the P2FA readme for additional documentation.

1. **Windows users**: Download and install Cygwin, a Bash shell for Windows. During this process you will be prompted to select the packages that you'd like to install. Under Python, select *python: Python language interpreter*. If you plan to build HTK from source, under Perl, select *perl: Larry Wall's Practical Extracting and Report Language*, and *perl_manpages: Perl manpages*. Under Devel, select *gcc-core*, and *make*. Under Utils, select *diffutils*. Finally, under Libs, select *zlib*.

   ***nix users:** You will need to have Python installed to run P2FA. Python 2.7 is best, however 2.6 works fine as well.)

2. Download the *nix source code (even if you're on Windows) for version 3.4 (*not* 3.41 or anything later) of the Hidden Markov Model Toolkit (HTK). You will first have to create an account, but it's free. (Note: Compiling a program from source can be a difficult process if you've never done it before. If using Windows, see the install bundle on the ORCA for pre-compiled binaries and skip to step 3.)

3. Extract the .tar file containing the HTK source, open Cygwin/Terminal, CD to the HTK directory, and enter `./configure`. Once the terminal is finished checking the source, enter `make all`, then `sudo make install`. (Note: if you're using OSX 10.5 or later, HTK 3.4.0 won't compile using the default commands, use these instead: `./configure --build=i686-apple-macos` and `make all CFLAGS=-DARCH='\"darwin\"'" -I/usr/include/malloc -I$PWD/HTKLib -I$PWD/HLMLib -I. -I.. -L/usr/X11/lib"`, then `sudo make install` as usual) On Windows, once finished, you should have 34 binaries in the HLMTools and HTKTools subdirectories in the HTK folder. Collect these and put them wherever you'd like (I created a folder called "HTK" in Program Files).

4. Download Sound eXchange (SoX), the latest version, and extract the files to a folder of your choosing (*nix users should download the source and compile/install it). (Note: P2FA requires the feature "polyphase", which was deprecated starting with 14.1.0 and removed in 14.3.0, however we're using a modified version of align.py which uses the replacement feature "rate" and reduces resampling times considerably. If you'd like to use the original script with "polyphase", download version 14.0.1 instead.)

5. Download P2FA and extract files to a folder of your choosing (Note: If using Windows, make sure the path doesn't have a space in it, e.g. "Program Files"). Download the modified align.py from the ORCA and replace the align.py found in your P2FA folder.

6. **Windows users:** Add HTK, and SoX to your path variables. To do so, open *System Properties* (from the Control Panel or Computer), clicking to the *Advanced* tab, then clicking the *Environment Variables* button near the bottom. Under the *System Variables* box, locate the *Path* variable, then click *Edit*. There will be a long list of paths here, so add a ";" separator to

the end of the list, then add the paths from HTK, and SoX, separating each with a ";". For example, I added this: *; C:\Program Files (x86)\HTK;C:\Program Files (x86)\Sox*

***nix users:** As long as you executed `make install` as superuser, the binaries should already be accessible by the system.

7. Download the most current dict file from the ORCA (which has had PNWEII words added to it), and place it in p2fa/model.

8. Now you should be ready to run P2FA. Do so by opening Cygwin/Terminal and executing the command `python align.py [options] soundfile.wav transcript.txt textgrid.TextGrid`, from within the P2FA directory or otherwise. After executing the script, you should receive a message informing you that the .wav file is being downsampled, and a while later the script should finish, generating a .TextGrid file.

   Important: the `align.py` command can only be issued from the folder containing the python script. If you have your files inside another folder (e.g., a processing file within the p2fa folder), run the script from the p2fa folder and specify the subfolder name in the path.

III. INTERVIEW PROCEDURE

1. Prepare and take paper copies of demographic questionnaire and consent form for respondents for use in the event that there is no internet access at the field site.

2. Upon arrival, try to find a suitable (i.e. quiet, isolated) area to conduct the interview.

3. Begin setting up equipment (i.e. recorder, stand, hotspot if necessary, iPads).

   NOTE: When recording a dyadic conversation, set the microphone polar pattern to 120°. When recording a single speaker, leave it at 90°.

4. Access the PNWE study website (the Study site, not the Public Website):
   `http-//www.artsci.washington.edu/nwenglish/study`
   Enter "webparticipant" as the password, then present the subject with the online consent form.

5. Once subject has finished reading the consent form, prompt them to enter their subject ID, and click "submit."

   IMPORTANT: In the event that no internet connection can be established, obtain subject consent using the paper forms.

6. Start the recording, beginning with speaking the date and subject ID(s) into the recording (so the audiofile can be linked with any documentation referencing subject ID information), and recording an oral confirmation from the respondents(s) that they have consented to be recorded.

7. After finishing the interview, again confirm that the subject(s) are satisfied with the recordings, and don't wish to delete any material.

8. Issue respondent pay ($15) to each subject, along with receipts (keep carbon copies).

9. Leave contact information and website information with respondent(s).

IV. TRANSCRIPTION

Summary: The aim of transcription is to render the recordings searchable for phonetic events we can analyze. We do this by generating orthographic transcripts which are then converted to force-aligned phonetic transcriptions, and can be subjected to acoustic measurements.

A. Orthographic Transcription (ELAN):

Create 7 tiers in ELAN: (It isn't critical this order be precisely followed)

Tier Number	Tier Name	Description
Tier 1	<SID>	Speaker 1 orthography (from transcript)
Tier 2	<SID>	Speaker 2 orthography (from transcript)
Tier 3	Interviewer1	Interviewer orthography (from transcript)
Tier 4	Interviewer2	Interviewer orthography (from transcript), if applicable
Tier 5	Themes	Conversation themes (see conversation coding guide)
Tier 6	Metadata	Utterance level metadata
Tier 7	IPA1	Phonetic Transcription (Speaker 1) copied here after work completed in P2FA/Praat (see below)
Tier 8	IPA2	Phonetic Transcription (Speaker 2) copied here after work completed in P2FA/Praat (see below)

Refer here to ELAN/transcription documentation

B. Phonetic Transcription (P2FA)

1. Open the recording in Praat (use "Open Long File" if necessary)

2. Look up the timestamps for the task you'd like to analyze in PNWE soundfile info.xls, and verify that they are correct. If the timestamps have not been entered into the spreadsheet, or if there are errors, note the beginning and ending times (in seconds) and add/amend the spreadsheet.

3. If the task is a formal one, use the associated template transcript, then listen through and make additions/alterations as needed. If the task is unscripted, make sure you have a transcript of it ready (i.e. in plain text format, stripped of conversational codings, preferably using Strip-Scripts.r).

4. While listening to the recording, set all false starts, interviewer speech, conversational partner speech (if the recording is a conversation), or other loud non-linguistic noises to silence ("Edit" > "Set Selection to Zero"). This will make for a better force-alignment, and ensure that we only have tokens from the speaker we are intending to analyze. If there is persistent, periodic background noise on the recording, consider using a program like Audacity to filter it out (select a few seconds of background noise, click "Effect" > "Noise Removal" > "Get Noise Profile", then select the part of the recording you'd like to remove noise from, click back to "Noise Removal" and click "OK"). When doing any editing of the recording, make sure to save it as a copy.

Some criteria: Leave words that aren't in the task's script in the recording, remove those that are cut off, drowned out by loud noise, or otherwise problematic for analysis. If during a formal task the speaker stops and speaks to the interviewer or someone else, this should be removed, as it not necessarily in the same speech style as the task.

Important: When setting segments of audio to silence, insert "{SL}" into the transcript at the appropriate point(s), otherwise P2FA may attempt to force-align the words in the transcript to total silence.

5. Run P2FA as detailed above, using the option `-s start_time` and `-e end_time` to constrain the aligner to the task you want to align. When aligning multiple tasks you should generate multiple TextGrids. Keep in mind that long tasks may take a very, very long time to align (e.g. 30+ minutes), so be patient.

   Example: `python align.py -s 1758 -e 2835 YH47HM3H_clean.wav YH47HM3H_WL.txt YH47HM3H_WL.TextGrid`

   This runs align.py on the wav file YH47HM3H_clean.wav (a copy of the entire recording which has been cleaned up for P2FA), constraining it with the -s and -e options to only the wordlist task, using the transcript file YH47HM3H_WL.txt, generating the TextGrid YH47HM3H_WL.TextGrid.

   Note: CMUdict is quite large, but it doesn't contain everything. If you get an error message saying "____ not found, skipping", first make sure you're using the most recent dict file, and if so, open the dict file and add the word near the end before the line "!ENTER []". Base your entry on the format of the other entries in the dictionary, and make sure to use Arpabet and indicate stress for the phonetic transcription. After adding words to the dict file, be sure to upload the new version to the ORCA.

C. Themes

We use a set of utterance-level annotations to tag "themes" that occur in the unscripted, conversational recording sessions recorded as part of each PNWE interview. A separate document in the ORCA, entitled `Conversational Coding Guidelines.docx`, outlines and illustrates our coding scheme. The goal was to provide a simple scheme to allow for systematic detection, tagging, and search of a minimal, but broad range of topics of interest to researchers in a range of disciplines who work with textual or audio-recorded and transcribed data. These themes are the basis for a content analysis (Stemler, 2001), and are intended to be interoperable with OLAC standards for metadata representation.

V. PREPARING FOR ANALYSIS

A. Hand-correcting Force-Aligned TextGrids

Although P2FA is a great help in the phonetic transcription of our recordings, it is no substitute for manual transcription. Thus, before any automated acoustic analysis can be done, all P2FA-generated TextGrids must be hand-corrected by stepping through the transcript and manually adjusting the boundaries. The following is a set of criteria for correcting vowel boundaries. (Note: hand-correction is very time-consuming, so it is recommended that only the tokens of interest be hand-corrected until such time as we can have the TextGrids fully hand-corrected.)

Onset

Look for…

- The first glottal pulse with representation at all frequencies
- An increase in periodicity
- Changes in the shape of the waveform
- Changes in amplitude (usually an increase, depending on the context)

Offset

Look for…

- The last glottal pulse with representation at all frequencies
- A decrease in or cessation of periodicity
- Changes in the shape of the waveform
- Changes in amplitude (usually a decrease, depending on the context)

General Criteria

- Always place boundaries at zero crossings
- With tricky cases (e.g. liquids and glides), try gating (playing successively longer portions of the waveform) to determine where a segment begins or ends

While hand-correcting, note any potential problems (e.g. creaky voice, background noise, formant tracking errors, hyperarticulation, etc.) in the notes tier. Pay attention to Praat's pitch track. If Praat is having trouble finding f0, make sure that it appears to be finding the formants correctly. If not, note this as well. In serious cases, the token may have to be thrown out.

B. Praat

Before running any scripts, double-check that the formant settings listed in formant settings.txt work for the speaker, and that the formants are being detected correctly. If not, try changing the settings to find something that works better, and notate the change in the textfile.

VI. PHONETIC ANALYSIS

Summary: The aims of the two-tiered (auditory and acoustic) phonetic analysis are to: 1) locate phonetic events for acoustic measurement, 2) determine the formant structures and durations of these events, 3) use auditory analysis to highlight information relevant to the sociolinguistic analysis (e.g., whether a vowel quality is raised, lowered relative to a geographic norm for that quality), 4) to prepare data for across-speaker and within-speaker comparison (i.e., normalization). This section is a listing of the tools we are using to analyze the data, with descriptions and instructions for each.

Because measuring durations and formants in continuous speech material can be a real challenge, we recommend starting with measuring the word list data, then measuring the reading passage data and linguistic task data, then the conversational data last. Relevant guidance from NSF proposal:

> "For wordlist data, targeted forms consist of all vowels collected in (h)Vt and (h)Vd contexts. For conversational data, targeted forms consist of the same vowel qualities, but collected in stressed, CVC contexts (including words that match wordlist forms, where these occur). Because number of repetitions of conversational forms also cannot be controlled, the typical practice of analyzing only the first three repetitions of eligible words produced after a "warm-up period" of 5-10 minutes into the conversation will be followed." (p. C-12)

PNWE Arpabet Vowel Analyzer.praat (link)
This script (modeled on Mietta Lennes' collect_formant_data_from_files.praat available at http://www.helsinki.fi/~lennes/praat-scripts/ and distributed under the GNU General Public License, copyright 4/7/2003) is designed to be run on a set of soundfiles and P2FA-generated TextGrids. It extracts duration (in ms), timestamps (in s), and F0/F1/F2/F3 from all intervals containing Arpabet vowels, and extracts labels for corresponding word and preceding & following phones, along with any notes present in the TextGrid. The script can also be constrained to a user-defined set of words using the "targets" option, and can insert demographics information from an accompanying tab-delimited text file. Before writing the results file, words are lowercased and Arpabet is converted to Unicode IPA.

- Remember to add a / or \ (depending on the OS) to the end of all paths.
- Soundfiles and TextGrids must have identical names in order for the script to match them up.
- Because the output is a text file containing IPA characters, make sure Praat's text writing preferences are set to UTF-8 in Windows, or UTF-16 in OSX before running the script.
- If you intend to view the results file in Excel, in Windows you will have to open it from within Excel and import it specifying UTF-8 encoding in order for the IPA characters to display correctly. On OSX, you will first have to convert the file to UTF-16 Little Endian using a text editor, then import it into Excel.
- To use the "targets" option, create a tab-delimited text file containing a list of the words (case-sensitive) you'd like to extract, separated by newlines, with "word" as the column header.
- To use the demographics information option, create a tab-delimited text file containing the relevant information with the following columns: "Speaker", "Sex", "Ethnicity", "Generation", "Age", and "SEC".
- To extract speaker and task information from filenames, make sure files follow this naming convention: "speakerID-taskID". Task codes are CS, DM, LX, RP, WL.

pnwe_get_205080.praat (link)
This script will append interval headers to your database spreadsheet, should you choose. It will then take token ID ("word"), overall duration, onset and offset timestamps, followed by formant measures at vowel onset, 20%, 50%, 80% and offset time locations. Timestamps are logged for all intervals. The "move cursor" function causes the cursor to move through your visible selection onscreen so you can monitor the locations at which Praat is taking measures.

You can invoke this script at any time (i.e. either before or after running GetFormants). On most of our analysis computers, this script is set up to be invoked with the F12 shortcut key. If you wish for your data to be appended to the `pnwe_<yourspreadsheet>_clean.txt` spreadsheet, you should run this script immediately after invoking GetFormants for each token. If you want to save the data to a separate output file from the GetFormants output. In this case, the order in which you run scripts doesn't matter. You can designate a separate output file, and then copy and paste the columns of `pnwe_get_205080.praat` output into the `pnwe_<yourspreadsheet>_clean.txt` spreadsheet when finished with a speaker.[1]

Strip-Transcripts.r (link)
This script strips conversational transcriptions of a range of commonly-encountered annotations. It is structured for use with the PNWEI recordings that have been transcribed in Praat tiers (and thus includes both Praat header info and markup, and discourse mark-up <@>, <XX>, <H>, and finally, timestamps), as well as with PNWEII recordings that have been transcribed in ELAN.

Task-Stats.r (link)
This script is intended to assist in the analysis of the PNWE recordings by reading in transcript files from a given task and generating word frequency statistics and cross-task comparisons, based either on another set of transcripts or the formal task scripts subjects are asked to read.

phonR (link)
phonR, developed by our own Dan McCloy, is an R package for normalizing and plotting vowels. For instructions, see the manual.

Importing Data into FileMaker
Things to keep in mind in importing your analyses into FileMaker:

1. You begin by importing data into the "Vowel Measures" layout. Doing this should populate fields in "Formal Linguistic Tasks" layout (specifically, in the Conversation, RP and WL tabs)

2. You can separately import data into the "Speaker Information" layout

[1] Added by ABW 10/2/08

3. Data should be automatically interleaved between the two records. In other words, FileMaker should figure out which speaker records go with your new tokens, since the tokens will be imported with speaker ID codes.

4. You should be importing from a spreadsheet that has the combined data from the Akustyk analysis and the 20/50/80 analysis. The order of the columns in this spreadsheet isn't important, but all the columns from these two analyses need to be present.

5. The spreadsheet you're importing from must be comma delimited for FM to see the source fields properly.

6. When adding new records into the FM database, use the radio button "Import New Records" (in the window "Import Field Mapping"). If you're updating records you've already imported (e.g., correcting a mistake), use "Update matching records in found set". *Don't* use "Update existing records in found set", as that will delete all the records previously in the database (including your colleagues'!).

7. When you "Update matching records in found set", you need to define the matched fields. These should be "Akustyk pkey_Spkr", "task", and "timestamp".

'Import Field Mapping' Specifications in FM

vowel_pk	→	Akustyk pkey_V
vowel_ipa	→	VowelTokenASCII
subject	→	Akustyk pkey_Spkr
vowel_code	→	HWC
word	→	Word
phone_p	→	Prec phone
place_p	→	P place
manner_f	→	F manner
voice_f	→	F voicing
cat_1	→	task
cat_2	→	SoundsMonophthongal
cat_3	→	Word Number
notes	→	user cat 3 comments
stress	→	stress
duration	→	V duration_sec
begin_cursor	→	timestamp
overall F0	→	WordF0
begin_F1	→	F1onset
begin_F2	→	F2onset
begin_F3	→	F3onset
20_F1	→	F1twenty
20_F2	→	F2twenty
20_F3	→	F3twenty
50_F1	→	F1mid
50_F2	→	F2mid
50_F3	→	F3mid

80_F1	→	F1eighty
80_F2	→	F2eighty
80_F3	→	F3eighty
end_F1	→	F1offset
end_F2	→	F2offset
end_F3	→	F3offset

VII. POST-ANALYSIS TASKS

Summary: Following analysis, it is crucial that the data be checked to ensure that no errors occurred during the process. This is a multi-stage process, involving poring over the spreadsheets, graphing the data, and going back to the TextGrids.

A. Checking and Correcting F1 and F2 measurement

After running the vowel analyzer, you should have a large spreadsheet of data in front of you. Because extraction of these measures was done automatically, it is important to check them at this stage in order to weed out any spurious data. This was accomplished by auto-generating an excel column that would highlight outliers, which were then systematically checked, with corrections being made. We focused on 50% measures first, then came back and did 20 and 80%. Procedures are described below. Other things to look for are empty cells, or ones that contain an error.

It is helpful to highlight cells in need of correction in order to more easily locate them later on.
The details of the checking and correcting process are described as the following:

- Measurements of each token will be checked at designated timestamp of the sound file. For example, the 20% timestamp was determined by using the following formula: Start time + ((End time – Start time)/5)
- F1 and F2 measurements are crosschecked with previous measurements in "PNWEI-measures-allspeakers-2014.xlsx", by comparing each token's F1 and F2 values at the timestamp that is closest to the desired value. Label measurements that have mismatches between 20 Hertz to 100 Hertz. Label measurements that have more than 100 Hertz yellow.
- After each session, the sound file is uploaded and noted on ORCA database.

Locating tokens that have mismatched sound file and text grid:

A list of temporary sound files are created and uploaded to the ORCA data under "sound files", future researcher could use these sound files to resolve mismatches between tokens and text grid, or mismatches between recorded timestamps and actual token location on Praat. To manually locate each vowel token in the 1000ms segments, use the view function on Praat, by typing the time stamp for each token (For example, if the start time and end time for one of the /e/ tokens in a 3200ms long sound file are 300ms and 400ms, one could test the timestamp for each of the temporary 1000ms segments). It would also be helpful to have a general idea of each token's location in the recording by referring to the word list task.

Creating temporary sound files for each speaker (if needed):

In cases the sound files do not have a matched text grid with each target token, due to previous merge of multiple smaller sound files, and temporary sound files were lost. Temporary sound files are created by splitting the mismatched sound files into 1000ms segments (when splitting the sound files on Praat, uncheck the "preserve timestamp" option).

B. Graphing the Data

Following an inspection of the raw data, the data should be graphed and inspected again. This allows for the easier isolation of outliers. The data should be graphed with 2 standard deviation ellipses, with the following precautions:

- only use CVC for /o, ɔ/ (i.e. don't plot "hall", or other words where retraction may bias the center of the distribution)

- If historic wordclass affiliation is not clearly agreed upon in the literature, follow the Labov symbol correspondences (to be consistent within the PNWE Project), and omit any questionable forms.

In generating plots, if you notice any unusual data points or clusters, make note of them and check them in the spreadsheet and the recording later.

C. Checking the Recording

When you've found some data that appears to be an error, make note of the timestamp, and use that to jump straight to that area of the waveform. Look for obvious signs of error, such as pitch halving/doubling, creaky voice, or background noise, and compensate if possible. If Praat is having trouble finding the formants for some other reason, try adjusting the formant settings and seeing if that improves things. If not, and if the formants are clearly visible, take the measurements manually and paste them back into the spreadsheet (Formant > Get frequency at cursor)

D. After Correcting a Cell

Make sure to always indicate in the spreadsheet when a value has been corrected. This should be done by highlighting the cell, and adding your name in the analyst column.

VIII. SOCIOLINGUISTIC ANALYSIS

Summary: The aims of the sociolinguistic analysis are to: 1) determine how speakers are using phonetic forms that may diagnose dialectal differences, 2) summarize data relevant to elucidating speaker's social backgrounds, 3) address dialect contact aims of the study.

A. FileMaker Pro Layouts

All information (demographic, sociolinguistic and phonetic) is gathered together in the FileMaker database. Each speaker has a "record", which contains several "layouts", across which that speaker's information is distributed:

1. Recording Information: Information about the audio recording(s) on which a speaker appears. (Enter data manually.)

2. Speaker Information: Human subjects, demographic and social network information, reliability data, Notes and Container fields for Cool data to post to public website. All information on these layouts (with the exception of underlined fields, which are imported from Akustyk spreadsheet) must be entered manually.

3. Speaker Summary Data Tables: Automatically-generated summary of formant values for vowel categories, by task. (Populates fields as you work. No data entry or script-running necessary.)

4. Linguistic Task Responses: manually input auditory transcriptions of formal linguistic tasks. (Can import the broad IPA transcription generated in Akustyk.)

5. Vowel measures: Receives individual vowel measure data; imported automatically from Praat spreadsheet. To input data, go to step 1, above.

6. Vowel codes and Linguistic task correspondences. These are word information and vowel code layouts with reference information about the data lists. For display purposes only. You don't enter data into these layouts.

   Note: For convenience, you can open the spreadsheet containing acoustic measures from several FileMaker views. Click the button, "Go to PNWE Spreadsheet." Note that you cannot view this data directly in FileMaker.

   Near future: ABW plans to create an import information button specific to each page.

B. Demographic data coding

At each interview, the researcher gathers information about respondents' demographics and social network. This information is elicited verbally and answers are recorded on the recording. Some info may be manually written in the respondents' information packets. At any time after a recording is made, the PI or REU can transfer the demographic and network information that's hand-recorded in the questionnaire packets into the FileMaker Pro database.

Notes: Some data is automatically coded in the phonetic analysis step by Akustyk (as Speaker metadata). These fields may be skipped. Before manually entering data,

a.) import **speaker** metadata from Akustyk file **`pnwe_spreadsheet_speaker.txt`** into `PNWEsociodemnet.fp7` layout **Speaker Information** (see 1-3, below, or alternatively, run automated script: Click button "Import Speaker Data")

b.) import **vowel** measures from Akustyk file **`pnwe_spreadsheet_clean.txt`** into `PNWEsociodemnet.fp7` layout **Vowel measures** (see 1-3, below, or alternatively, run automated script: Click button "Import Midpoint Measures", then separately click "Import Interval Measures")[2]

1. In your desired layout**,** press button **Import**.

2. You will see a dialogue box allowing you to select the Akustyk dbase file. Select **File.**

3. Complete Import Field Mapping. Match fields in Akustyk source file to fields in target file. You only need to do this once per layout (each layout "remembers" the last file it imported (and field associations). You will need to use information in Appendix E to see how the Akustyk Codes translate into various levels of the independent variables.

   Notes: Skip importing the Akustyk "age", "age group" fields—we'll enter actual data in FileMaker. **IMPORTANT:** Tell FileMaker to skip the first row (which contains headers).

4. Once speaker data are entered into the FileMaker database, each speaker will have one record, consisting of 5 layouts containing **all** information for that speaker.

5. For manual coding of demographic information in FileMaker, see Appendix F.

C. Generating Summary reports

We can use the "Speaker Summary Data Tables" layout to generate summary descriptive statistics using reports and frequency tables. Or, the summary information and fields may be exported in .txt format to another analysis program such as Excel.

Recoded vowel class (from Akustyk notation to Plotnik notation) using Excel's search-and-replace function:

	Akustyk		Plotnik
(ow)	11	→	62
(iy)	1	→	11
(i)	2	→	1
(e)	3	→	2
(ey)	10	→	21
(ae)	4	→	3
(ay)	27	→	41
(wedge)	9	→	6
(aw)	107	→	42

[2] Replaces advice to import from the **two** Akustyk files ...select.txt and ...interval.txt. We are importing from the ...clean.txt files which have the output of both the GetFormants and 205080 scripts. ABW 9/9/09.

(oy) 28 → 61
(uw) 7 → 72
(ah) 6 → 43
(ir,er,r) 105 → none (this isn't a vowel); I provisionally called these 44 (ahr), but delete them from set later.

*note: this specific order is important, to retain integrity of individual classes

APPENDIX A: AKUSTYK CODES

Regionality
Use this pulldown to register respondent's regionality. We will code R1, R2, R3, R4, R5+ as Akustyk "speaker category 1" (which gives only 5 options).

R1 (indigene speaker, raised locally, with indigenous parents)
R2 (indigene speaker, raised locally, 1 indigenous parent)
R2.5 (indigene speaker, raised locally, interloper parents)
R3 (indigene speaker, partly raised outside PNW during critical period, interloper parents)
R4 (indigene speaker, raised entirely outside PNW during critical period, interloper parents)
R5 (non-indigene, partly raised in PNW during critical period, indigenous parents)
R6 (non-indigene, raised entirely outside PNW, 1 indigenous parent)
R7 (non-indigene speaker, raised entirely outside PNW, with interloper parents)

Note: the judgement sample targets only indigenes, so it's ok that Akustyk gives us only 5 levels for variables. We wouldn't use the last two anyway. We want to use all 7 levels for the random sample and the larger study.

SEC
Use this pulldown to register respondent's self-reported socioeconomic class at birth, as Akustyk "speaker category 2". The questionnaire gives only 3 options (upper, middle and working), but sometimes speakers don't like our categories and use ones of their own. Put down EXACTLY what the speaker gave in their questionnaire.

Upper class: SECBirth_UC
Upper-middle class: SECBirth_UMC
Middle-class: SECBirth_MC
Lower-middle class: SECBirth_LMC
Working class: SECBirth_WC

Note: Occupation codes will be used to compute a more specific SEC_Now value automatically in FileMaker Pro database).

Year of Birth
Use this pulldown to register the age cohort for the respondent's year of birth (age and year of birth will be registered elsewhere), as Akustyk's "speaker category 3":

1900-1950
1951-1971
1976-1986
1987 and afterward
("none" should never be chosen. Akustyk just requires no field be left empty)

Correspondences for ordinal variables in Akustyk's named independent variable fields:

Name
Enter JSID or RSID code.

Sex
Select male or female.

Age
This will be calculated automatically in FileMaker. Do not complete field.

Age Group
Leave blank. Speaker Category 3 encodes this information as YrofBirth_range

Ethnicity
We code for 4 levels in the larger study:

1-Native American
2-Asian-American
3-African-American
4-Caucasian

SES (=SEC)
Use this pulldown to register respondent's self-reported socioeconomic class NOW. The system below matches that used in SECatBirth category, above. The questionnaire gives only 3 options (upper, middle and working), but sometimes speakers don't like our categories and use ones of their own. Put down EXACTLY what the speaker gave in their questionnaire.

1-Upper class
2-Upper-middle class
3-Middle-class
4-Lower-middle class
5-Working class

Note: Occupation codes will be used to compute a more specific SEC_Now value automatically in FileMaker Pro database).

Education
We code for 4 levels:

1-Elementary
2-Jr. High/High/GED
3-B.A./B.S.

4-M.A./Ph.D.

Occupation
Leave blank. Occupation name will be specified in FileMaker Pro.

Network
Leave blank. We use a detailed procedure for calculating network strength in FileMaker Pro.

Neighborhood
We code for two levels:

1 - mobile
2 - established (historic)

Comment Key (Added 10/1/08 by Rob Squizzero[3])
When entering comments in this section (e.g. raised, creaky voicing, etc.), if there is more than one comment, be sure to separate them with an underscore rather than a comma (e.g. F_fronted). Inserting a comma will distort alignment in the spreadsheet and in File Maker Pro.

Also note the following key for comments (Note: "omit" categories, to be removed before analysis):

F – omit: formant reading problem (Praat error)
R – repeat
C – corrected
Cr – omit?: creaky voicing
Mp – omit: mispronunciation
Retracted – retracted
Br – breathy voicing
Wh – whisper
Raised – raised
Lowered – lowered
Fronted – fronted
Reduced – vowel sounds shortened or nearly deleted (ABW)
P – Pitch-reading problem (MS)
Vdeleted – omit: token word sounds vowel-less (ABW)

[3] Rob's comments are relative to his own dialect of English – Providence, RI, which has vowels which show very little deviation from the positions of cardinal vowels on an IPA chart. Rob assumed /ae/ raises before nasals and marked as lowered if not raised.

APPENDIX B: SUBJECT ID CODES

Overall coding format for Seattle-area respondents

1 2 3 4 5 6 7
City - Neighb - Spkr# - Ethnicity - M/F - Generation - Family
(e.g., [SY9AF1A])

In the random sample: first letter in code will be "R" for "random sample", in slot occupied by "S" that appears in Seattle speaker's codes as in SY9AF1A example above; second letter should be city of "random sample" speaker.

1 City:
S - Seattle
P - Portland

2 Neighborhoods:
A - Bryant
B - Ballard
BI - Bainbridge Island
BL - Bitter Lake
BO - Bothell
C - Central District/South Seattle
D - Normandy Park
E - West Seattle
G - Green Lake
H - Beacon Hill
HH - Hawthorne Hill
I - Phinney Ridge
K - Kirkland
L - Lake Forest Park
M - Skykomish
N - Northgate
O - Shoreline
P - Sand Point
Q - Queen Anne Hill
R - Ravenna
S - Seward Park
T - Yesler Terrace
U - U-District
V - View Ridge
W - Wedgwood
Y - Puyallup

3 Speaker Number:
Continuous from 1 - total # in sample

4 Ethnicity:

C - Caucasian
S - Asian-American
A - African-American
H - Hispanic-American
N - Native-American

5 Gender:
M/F

6 Generation (by Date of Birth):
1 - 1900-1950
2 - 1951-1976
3 – 1976 to age 18 at time of study[4]

7 Family:
A-Z (code does not correspond to family name in any way)

Others:
For spkr# and family coding, see spreadsheet "PNWE soundfile info" for tallies so far. In Random Sample Recording Info Notes: W = White pages recruitment method; R = Random number generator recruitment method

Overall coding format for Eastern WA respondents
1 2 3 4 5 6 7
Region - City/Town - Spkr# - Ethnicity - M/F - Generation - Family
(e.g., [YY42HF2A])

1 Region:
Y - Yakima Valley

2 Cities/Towns:
G - Grandview
H - Harrah
M - Mabton
R - Granger
S - White Swan
T - Toppenish
U - Sunnyside
W - Wapato
Y - Yakima
Z - Zillah

7 Family:
NB: Family code restarted at A for YV sample.

[4] Lowest cutoff used to be 1986 (till second phase of the study highlighted fact that we were cutting off above the minimum age HSD would allow, forcing exclusion of eligible subjects). Sale (1976) Seattle history phases: 1800-50; 1850-1900; 1900-50; 1950-present.
Last revision: 10/22/18, 6:10 PM

Appendix C: Friend Network Coding

Neighborhood

We code for 7 different general regions:

NE – north-end
SE – south-end
ES – eastside
WS – West Seattle
NS – suburbs to the north
SS – suburbs to the south
O - other

Ethnicity

C – Caucasian
A – African-American
S – Asian-American
N – Native American
M - Mixed
O - Other

Gender

M – Male
F – Female

(e.g., NECM is north-end, Caucasian, male)