Lecture 18: Working with Speech Corpora, Datasets

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

▶ **Speech data**

- Speech corpora, datasets
- PRAAT
  - Python libraries for textgrid, praat
- Command-line exploration
  - Conversion tools: sox, ffmpeg

Well-known speech datasets, corpora

▸ **Buckeye Corpus** (Pitt et al. 2005)
  - ◆ Python interface! https://github.com/scjs/buckeye/blob/master/Quickstart.ipynb

▸ **TIMIT** (Garofolo et al. 1993)
  - ◆ 10 sentences read by 630 speakers from 10 US dialect regions
  - ◆ Orthographic transcription and phonetic annotation

▸ **Switchboard corpus** (Godfrey et al. 1993, 1997)
  - ◆ Phone conversations between strangers on assigned topic, 2400 conversations by 543 speakers, many US dialects represented

▸ **TalkBank** project (MacWhinney, at CMU!)
  - ◆ Multiple research focus areas: L1 acquisition, multilingualism, etc.
  - ◆ Data contributed by many researchers
  - ◆ Known for CHAT format and CLAN program

Also: APLS at Pitt!

▸ **CORAAL** (Corpus of Regional African American Language)
  - ◆ Recorded speech from regional varieties of AAL, includes audio recordings along with time-aligned orthographic transcription, all downloadable

What do *linguists* do with speech data?

▸ Measuring duration: VOT (Voice Onset Time), etc.

▸ Measuring formants, F0/pitch

▸ Measuring amplitude, frequency

▸ Audio format conversion

 ◆ WAV, MP3, FLAC

 ◆ Channels, sampling rates, etc.

▸ Edit and manipulate sound

 ◆ Crop, copy, slice, paste…

 ◆ Manipulate pitch, duration…

What tool do we
use for these,
I wonder…?

PRAAT

https://www.fon.hum.uva.nl/praat/
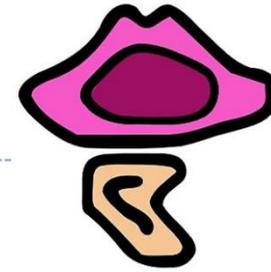
- ◆ Everyone's favorite phonetics data analysis tool
- ◆ Powerful, versatile, venerable... and idiosyncratic

▶ Logo change was very much celebrated (or not...):

- ◆ https://blogs.umass.edu/linguist/2020/10/19/umass-redesign-of-praat-logo/
- ◆ Looks like they changed back!

▶ Using Praat for Linguistic Research, by Will Styler:

- ◆ https://wstyler.ucsd.edu/praat/

▶ Paat Scripting Tutorial, by Eleanor Chodroff:

- ◆ https://eleanorchodroff.com/tutorial/PraatScripting.pdf

Praat + TIMIT
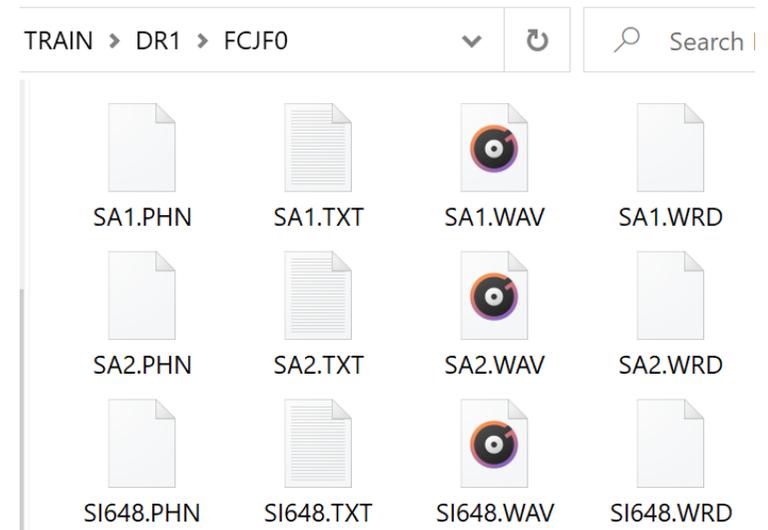
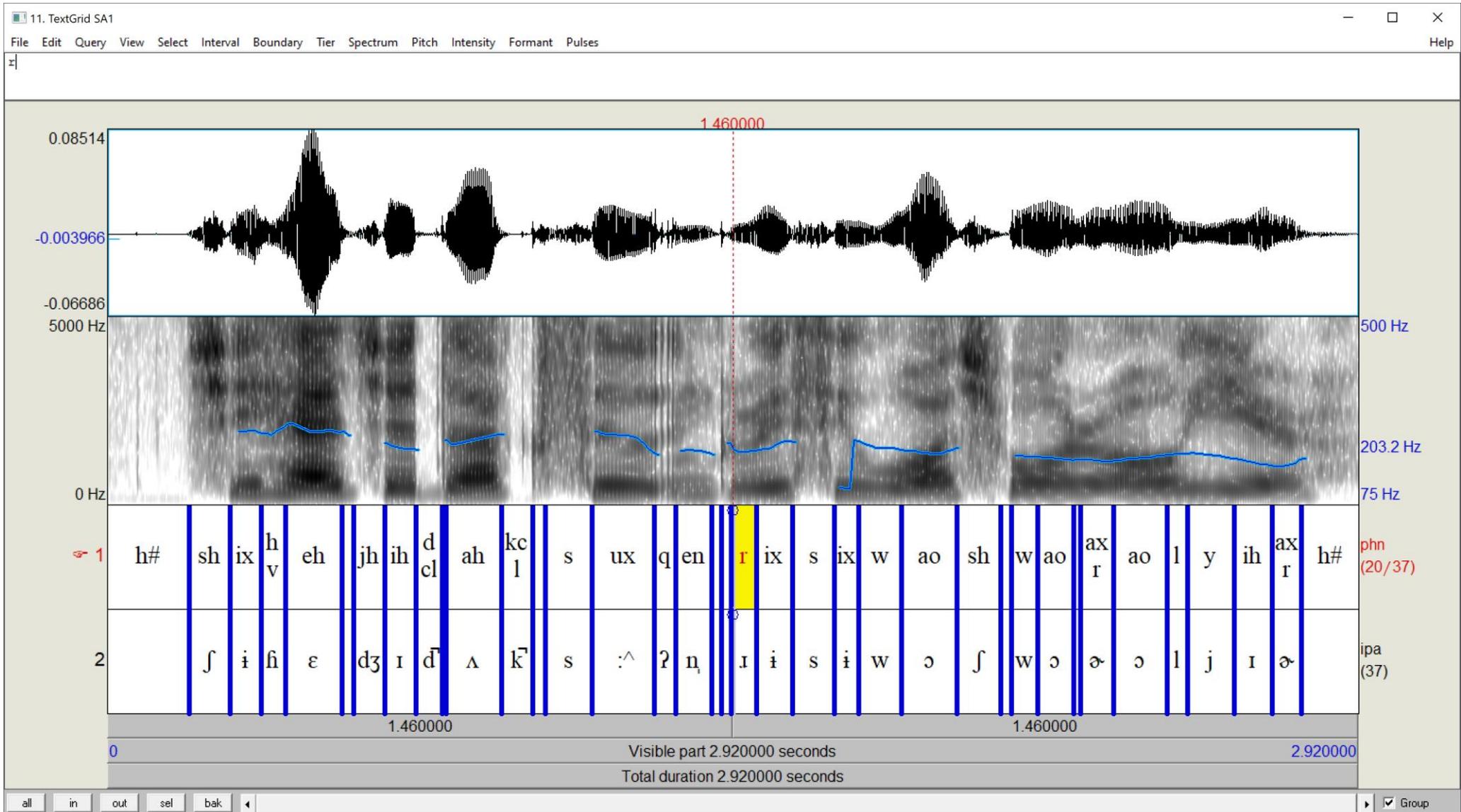▶ **An excerpt of TIMIT dataset is available on our GitHub org, in "Licensed-Datasets"**

- ◆ Get it by pulling from the repo.

▶ **Many of you have Praat on your laptop already**

- ◆ Open up "SA1.*" files in Praat, explore, see what you can do!

- ◆ Also: command-line exploration

Open .WAV file first, and then the rest after

Your will get warnings with some txt files

TRAIN > DR1 > FCJF0			
SA1.PHN	SA1.TXT	SA1.WAV	SA1.WRD
SA2.PHN	SA2.TXT	SA2.WAV	SA2.WRD
SI648.PHN	SI648.TXT	SI648.WAV	SI648.WRD

TIMIT data in Praat

TIMIT data in command line

```
narae@T480s MINGW64 ~/Desktop/speech/TRAIN-DR1-FCJF0
$ ls
SA1.PHN   SA2.PHN   SI1027.PHN   SI1657.PHN   SI648.PHN   SX127.PHN   SX217.PHN   SX307.PHN   SX37.PHN   SX397.PHN
SA1.TXT   SA2.TXT   SI1027.TXT   SI1657.TXT   SI648.TXT   SX127.TXT   SX217.TXT   SX307.TXT   SX37.TXT   SX397.TXT
SA1.WAV   SA2.WAV   SI1027.WAV   SI1657.WAV   SI648.WAV   SX127.WAV   SX217.WAV   SX307.WAV   SX37.WAV   SX397.WAV
SA1.WRD   SA2.WRD   SI1027.WRD   SI1657.WRD   SI648.WRD   SX127.WRD   SX217.WRD   SX307.WRD   SX37.WRD   SX397.WRD

narae@T480s MINGW64 ~/Desktop/speech/TRAIN-DR1-FCJF0
$ cat *TXT
0 46797 She had your dark suit in greasy wash water all year.
0 34509 Don't ask me to carry an oily rag like that.
0 49460 Even then, if she took one step forward he could catch her.
0 45466 Or borrow some money from someone and go home by bus?
0 57856 A sailboat may have a bone in her teeth one minute and lie becalmed the next.
0 24679 The emperor had a mean temper.
0 27751 How permanent are their records?
0 23143 The meeting is now adjourned.
0 36250 Critical equipment needs proper maintenance.
0 39220 Tim takes Sheila to see movies twice a week.

narae@T480s MINGW64 ~/Desktop/speech/TRAIN-DR1-FCJF0
$ head SA1.PHN
0 3050 h#
3050 4559 sh
4559 5723 ix
5723 6642 hv
6642 8772 eh
8772 9190 dcl
9190 10337 jh
10337 11517 ih
11517 12500 dcl
12500 12640 d

narae@T480s MINGW64 ~/Desktop/speech/TRAIN-DR1-FCJF0
$ head SA1.WRD
3050 5723 she
5723 10337 had
9190 11517 your
11517 16334 dark
16334 21199 suit
21199 22560 in
22560 28064 greasy
28064 33360 wash
33754 37556 water
37556 40313 all

narae@T480s MINGW64 ~/Desktop/speech/TRAIN-DR1-FCJF0
```

Utterance tier

Phone tier

Word tier

▸ Use cat, less, grep!

Essentially "Quick brown fox..." sentences for English speech sounds

TIMIT data in command line

```
Jane Eyre@T480s MINGW64 ~/Documents/Data_Science/Lice
s Speech Corpus/timit/TIMIT/TRAIN/DR1/FCJF0 (main)
$ grep dh *PHN
SA2.PHN:29000 29490 dh
SI648.PHN:27613 28841 dh
SI648.PHN:46640 46990 dh
SX127.PHN:2231 2834 dh
SX217.PHN:13785 14590 dh
SX307.PHN:1960 2170 dh

Jane Eyre@T480s MINGW64 ~/Documents/Data_Science/Lice
s Speech Corpus/timit/TIMIT/TRAIN/DR1/FCJF0 (main)
$ cat SA2.TXT
0 34509 Don't ask me to carry an oily rag like that.

Jane Eyre@T480s MINGW64 ~/Documents/Data_Science/Lice
s Speech Corpus/timit/TIMIT/TRAIN/DR1/FCJF0 (main)
$ grep ae *PHN
SA2.PHN:4600 6864 ae
SA2.PHN:22266 24898 ae
SA2.PHN:29490 32279 ae
SI1027.PHN:41210 43040 ae
SI648.PHN:12040 13800 ae
SX127.PHN:10160 11640 ae
```

Which files have /ð/ sound?

▸ Use cat, less, grep!

TextGrid

- ▸ Praat was able to parse TIMIT's PHN file format (phone tier)

- ▸ Saving it out to a proper TextGrid file →

- ▸ However, Praat couldn't handle:
  - ◆ SA1.TXT (utterance tier)
  - ◆ SA1.WRD (word tier)
  - ← How to get them into TextGrid?

  There are python libraries for that!

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 2.92
tiers? <exists>
size = 2
item []:
    item [1]:
        class = "IntervalTier"
        name = "phn"
        xmin = 0
        xmax = 2.92
        intervals: size = 37
        intervals [1]:
            xmin = 0
            xmax = 0.19062500000000002
            text = "h#"
        intervals [2]:
            xmin = 0.19062500000000002
            xmax = 0.2849375
            text = "sh"
        intervals [3]:
            xmin = 0.2849375
            xmax = 0.3576875
            text = "ix"
        intervals [4]:
            xmin = 0.3576875
            xmax = 0.415125
            text = "hv"
        intervals [5]:
            xmin = 0.415125
            xmax = 0.54825
            text = "eh"
        intervals [6]:
```

TextGrid and Praat in Python: libraries

▶ There are many Python libraries that let you manipulate TextGrid data format. Some even directly interface with Praat application.

▶ **textgrid**

- ◆ https://github.com/kylebgorman/textgrid
- ◆ Simple functionality for textgrid format manipulation

▶ **praat-textgrids**

- ◆ https://github.com/Legisign/Praat-textgrids

▶ **parselmouth**

- ◆ https://github.com/YannickJadoul/Parselmouth
- ◆ Extensive functionality, directly accessing Praat's code, provides Python interface

.WAV format?

▶ Even though PRAAT was able to open the .WAV files, Windows 10 cannot…

▶ These files are not really .WAV…

◆ **SPHERE format**, normally with .SPH extension.

▶ How to convert to WAV?

SA1.WAV

Can't play

This file isn't playable. That might be because the file type is unsupported, the file extension is incorrect, or the file is corrupt.

0xc00d36c4

Send feedback

Close
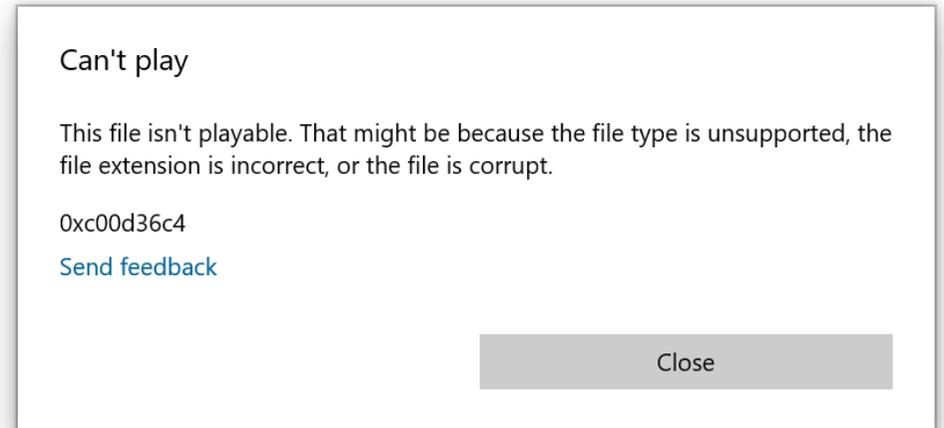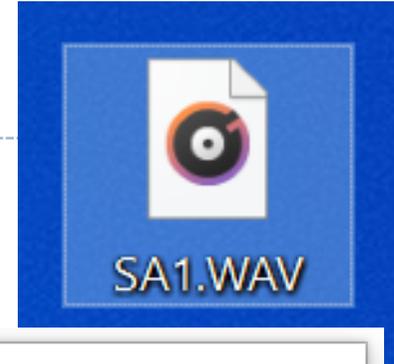
Solution 1: Praat script

▶ Write a praat script

♦ ([Or, grab someone else's...](#))

```
# prep_audio_mfa.praat
# Written by E. Chodroff
# Oct 23 2018
# extract left channel and resample to 16 kHz for all wav files in a directory

### CHANGE ME!
# don't forget the slash at the end of the path
dir$ = "/Users/Eleanor/Desktop/align_input/"
###

Create Strings as file list: "files", dir$ + "*.wav"
nFiles = Get number of strings

for i from 1 to nFiles
        # read in WAV file
        selectObject: "Strings files"
        filename$ = Get string: i
        Read from file: dir$ + filename$

        # extract left channel
        Extract one channel: 1

        # resample to 16kHz with 50 point precision (default)
        Resample: 16000, 50

        # save WAV file
        Save as WAV file: dir$ + filename$

        # clean up
        select all
        minusObject: "Strings files"
        Remove
endfor
```

Solution 2: SoX + bash shell

▶ SoX: Sound eXchage

  ◆ https://sourceforge.net/projects/sox/

```
sox <input-file> -b 16 -t wav <output-file>
```

```
narae@T480s MINGW64 ~/Desktop/FCJF0
$ alias sox="/d/util/sox-14.4.2/sox.exe"

narae@T480s MINGW64 ~/Desktop/FCJF0
$ ls
SA1.PHN    SA2.WAV      SI1657.PHN    SI648.WAV    SX217.PHN    SX307.WAV    SX397.PHN
SA1.TXT    SA2.WRD      SI1657.TXT    SI648.WRD    SX217.TXT    SX307.WRD    SX397.TXT
SA1.WAV    SI1027.PHN   SI1657.WAV    SX127.PHN    SX217.WAV    SX37.PHN     SX397.WAV
SA1.WRD    SI1027.TXT   SI1657.WRD    SX127.TXT    SX217.WRD    SX37.TXT     SX397.WRD
SA2.PHN    SI1027.WAV   SI648.PHN     SX127.WAV    SX307.PHN    SX37.WAV     true_wav/
SA2.TXT    SI1027.WRD   SI648.TXT     SX127.WRD    SX307.TXT    SX37.WRD

narae@T480s MINGW64 ~/Desktop/FCJF0
$ sox SA1.WAV -b 16 -t wav true_wav/SA1.wav

narae@T480s MINGW64 ~/Desktop/FCJF0
$ ls true_wav/
SA1.wav
```

converting a single file

Solution 2: SoX + bash shell

```
narae@T480s MINGW64 ~/Desktop/FCJF0
$ for x in *WAV
> do
> sox $x -b 16 -t wav true_wav/$x
> echo $x finished
> done
SA1.WAV finished
SA2.WAV finished
SI1027.WAV finished
SI1657.WAV finished
SI648.WAV finished
SX127.WAV finished
SX217.WAV finished
SX307.WAV finished
SX37.WAV finished
SX397.WAV finished

narae@T480s MINGW64 ~/Desktop/FCJF0
$ ls true_wav/
SA1.wav   SI1027.WAV   SI648.WAV   SX217.WAV   SX37.WAV
SA2.WAV   SI1657.WAV   SX127.WAV   SX307.WAV   SX397.WAV
```

for loop in bash!

```
for x in *.WAV
do
sox $x -b 16 -t wav true_wav/$x
echo $x finished
done
```

Declared as x, subsequent references as $x

Command-line conversion to mp3: with ffmpeg

▸ `ffmpeg -i input.wav output.mp3`

Again, how to do this with *every* wav file…?

For loop in bash

```
for x in *.WAV
do
fname=`basename $x .WAV`
newname=$fname.mp3
echo $newname
ffmpeg -i $x converted/$newname
done
```

x is newly created as a variable

$x: subsequent mention requires $ prefix

Note use of backticks `

New file names with .mp3 extension

Loop content begins with do and ends with done

General-purpose audio/video manipulation software

▶ [Audacity](#)

  ◆ Open-source audio software

▶ [SoX](#)

  ◆ Sound eXchange; audio format conversion tool ◀┈┈┈┈┈

▶ [FFmpeg](#)

  ◆ For recording and converting audio/video data ◀┈┈┈┈

> Powerful command-line tools!!

> https://musicinformationretrieval.com/sox_and_ffmpeg.html

Wrapping up

- Next class:
  - Forced alignment

- Project presentations coming up.