Lecture 19: Speech Data

LING 1340/2340: Data Science for Linguists

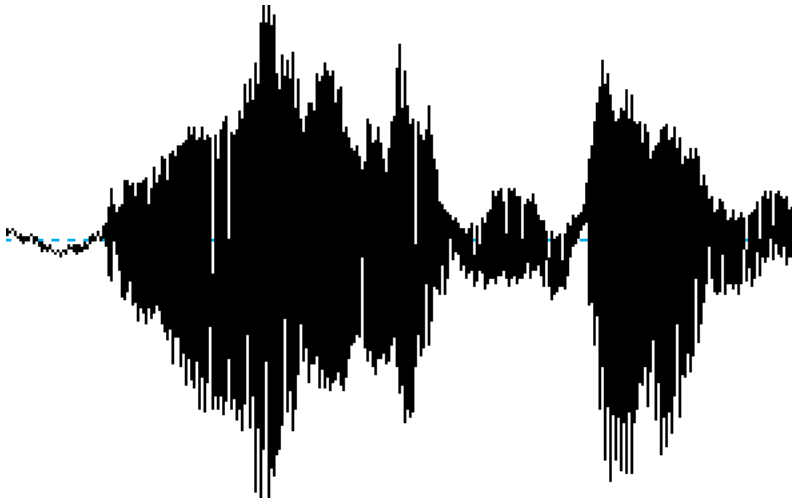
Na-Rae Han

Objectives

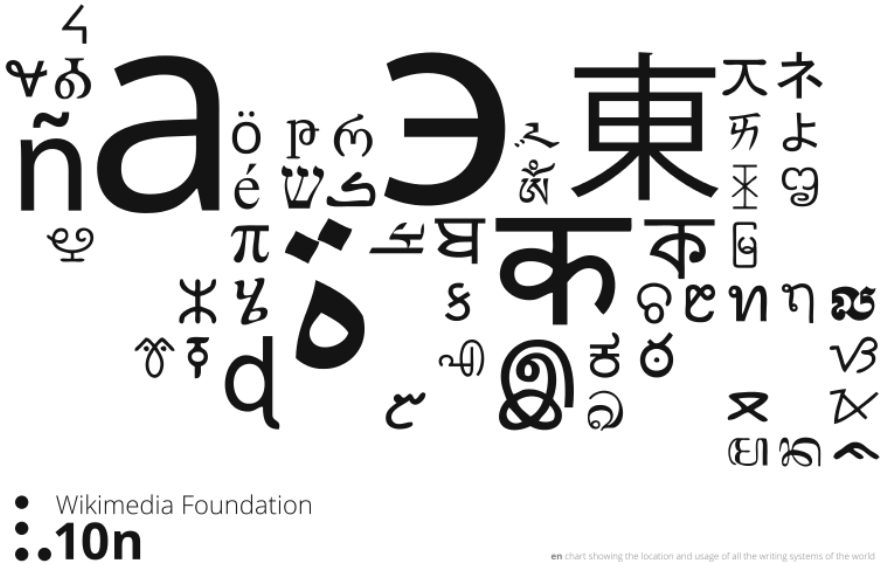
- Speech data
  - Speech corpora, datasets
  - PRAAT
  - Command-line exploration

Speech vs. Writing

- Ubiquitous to human communities
- Spontaneous
- Humans acquire speech without instruction



- Invented, many communities without
- Deliberate
- Requires instruction to learn



What to do with speech data?

- Analyze it directly.
  - Language identification
  - Phonetic research
  - Informing models (example below)
- Convert it to text, then text-process for downstream tasks
  - ASR (Automatic Speech Recognition) and ASU (… Understanding)
  - Automatic closed-captioning
- The other direction:
  - Speech Synthesis / Text-to-Speech (TTS)
  - Conversational Agents

Speech sounds: how to encode/represent?

- IPA, ɒbvɪəslɪ…
  - But IPA chars are Unicode characters, difficult to use directly

- Do you remember CMU Pronouncing Dictionary?

```
>>> from nltk.corpus import cmudict
>>> prondict = cmudict.dict()
>>> prondict['anxious']
[['AE1', 'NG', 'K', 'SH', 'AH0', 'S'], ['AE1', 'NG', 'SH', 'AH0', 'S']]
>>>
```

  - Uses **ARPABET**: https://en.wikipedia.org/wiki/ARPABET
    - ASCII-based representation of English speech sounds
  - CMU pronouncing dict is used in all sorts of English speech technologies…
  - Also: https://heardle.glitch.me/

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 corpora (MacWhinney, at CMU!)
  - Multiple research focus areas: L1 acquisition, multilingualism, etc.
  - Data contributed by many researchers
- 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
- Venerable, powerful, versatile… and idiosyncratic

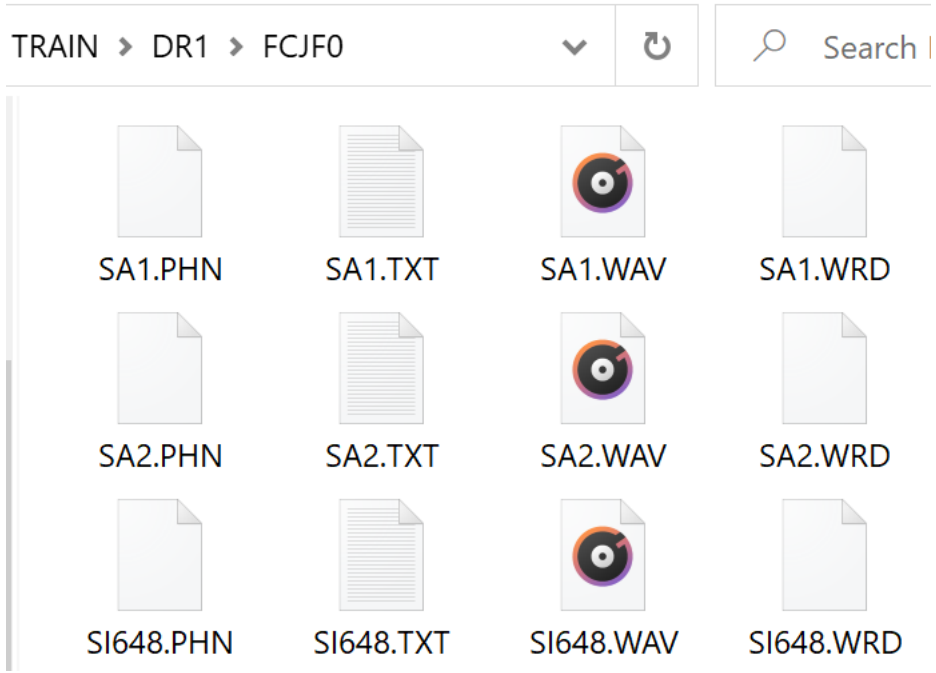
- Logo change was very much celebrated (or not…):
  - https://blogs.umass.edu/linguist/2020/10/19/umass-redesign-of-praat-logo/

- 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

Activity
7 minutes

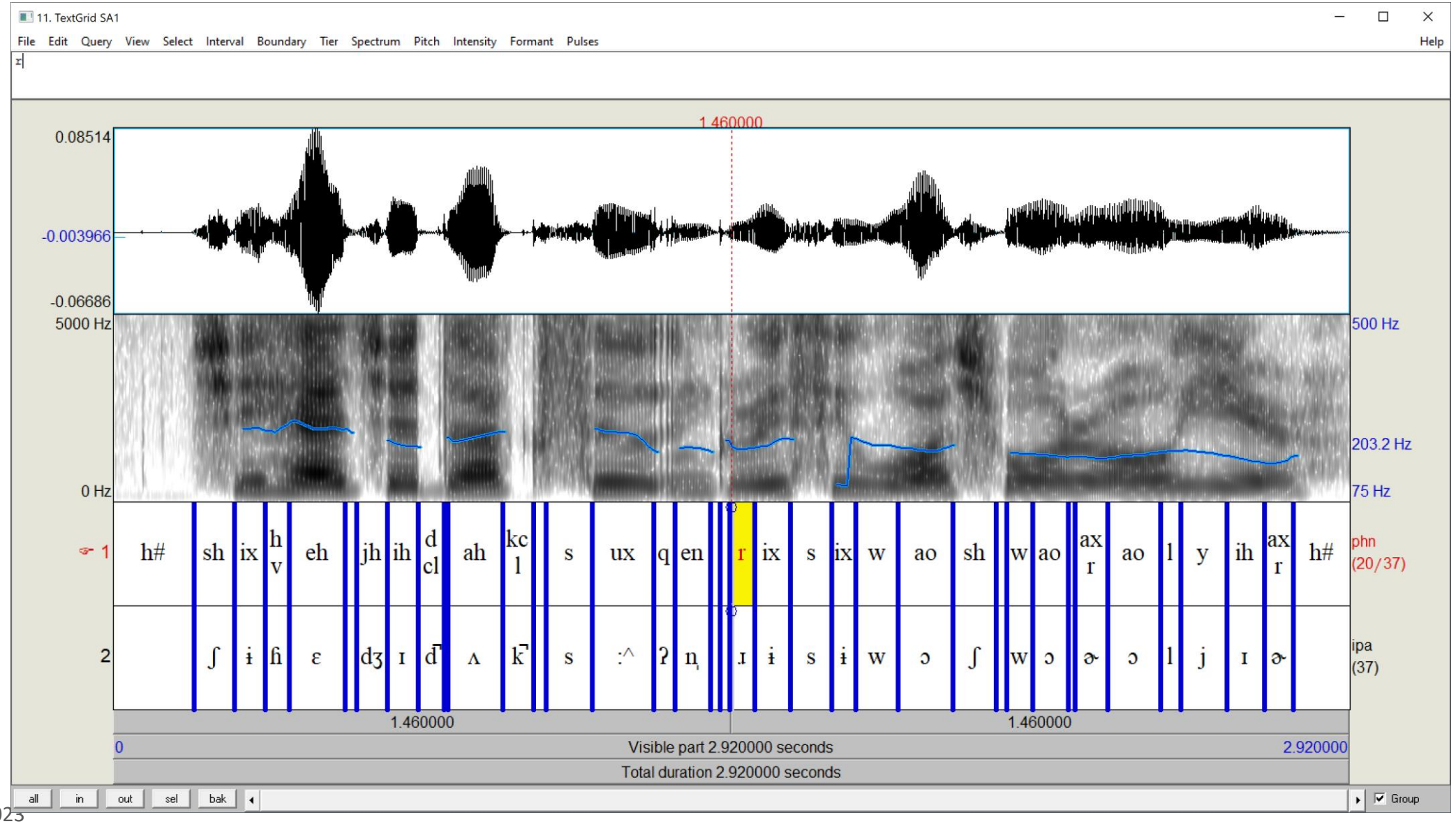
- An excerpt of TIMIT dataset is available on our GitHub org, in "Licensed-Datasets"
  - Get it by pulling from the repo.
- Many of you have Praat on your laptop already
  - Pair up, open up "SA1.*" files in Praat, explore, see what you can do!
  - Also: command-line exploration



Open .WAV file first, and then the rest after

Your will get warnings with some txt files

TIMIT data in Praat



```
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

TIMIT data in command-line

- Use `cat`, `less`, `grep`!

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

Wrapping up

- Next class:
  - More on praat
  - Command-line conversion
  - Forced alignment overview
- 3rd progress report due Monday!
- Also coming up: project presentations. Dates/presenters fixed… check yours!