Corpus Phonetics

Eleanor Chodroff
University of York

Newcastle University 11 July 2019

Please download CorpusPhonetics.zip ☺

https://eleanorchodroff.com/CorpusPhonetics.zip

It’s big.

We’ll be adding to it and there may be a few more files tomorrow.

On phonetics:

From 1966 on, there has been “surprisingly little change in style and scale of research”

Mark Liberman
Talk on “A New Golden Age of Phonetics?” at JHU CLSP 2011

On phonetics:

From 1966 on, there has been “surprisingly little change in style and scale of research”

Mark Liberman
Talk on “A New Golden Age of Phonetics?” at JHU CLSP 2011

Style of phonetic research

Hand segmentation and labelling of speech data
Manual phonetic measurement

< TIME-CONSUMING, TEDIOUS >

Scale of phonetic research: small!

But the computer!

Report in 1966 from the Automatic Language Processing Advisory Committee in the National Academy of Sciences acknowledging the power of the computer, especially for linguistic analysis

Enter computational linguistics and a “new science of language”

Computational linguistics embraced in engineering

(NLP, ASR systems, language in technology, etc.)

But have we actually achieved a new *science* of language?

It’s not like we haven’t had decent computational power for awhile

Computers in 1960s
Supercomputers in 1980s
Laptops now ubiquitous (even back in 2010)

2010 was almost 10 years ago!

And we do use comp power in speech science:
Digital audio
Generating spectrograms (thank goodness)
Phonetic measurements
Statistical analysis

But even in 2019, can we say we have a new science of language?

Are large-scale and automated analyses of speech data actually commonplace?

Well, what counts as a large-scale and automated analysis?

(Enter gray area)

Increased availability of and access to:

large spoken corpora

speech processing tools

Increased availability of and access to:

large spoken corpora

speech processing tools

Enabled by:

More powerful computers (and again, their availability and accessibility)

Advances in speech engineering (ASR)

Can you help a scientist out?

Large spoken corpora

Data available through:
LDC
ELRA
Online

Many collected by government defence agencies (DARPA, MoD)
Increased interest in collecting large corpora

Speech processing tools

Praat: phonetic measurement
Praat scripting: automating phonetic measurement

Forced alignment: automates segmentation

“Refined alignment”: more precise automatic alignment of a particular segment or set of segments (example: AutoVOT)

Corpus phonetics

Using automated approaches to process and analyse speech data

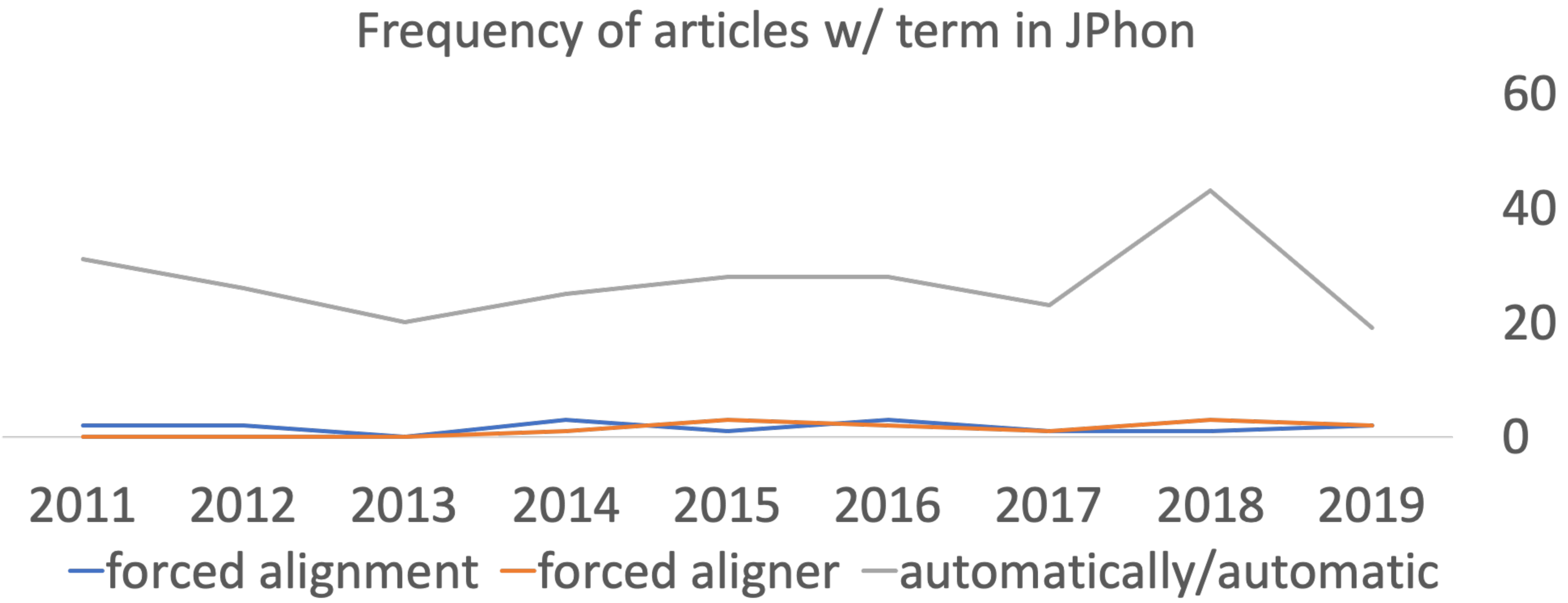
Generally corresponds to the ability to process large quantities of data, though large scale is not a requirement

Aspects of data preparation, processing, or analysis are automated

But even in 2019, can we say we have a new science of language?

Are large-scale and automated analyses of speech data actually commonplace?

Quick search of Journal of Phonetics



We have the tools, we have the data

Why isn’t corpus phonetics more commonplace?

Maybe we need a little more training

YAY, workshop!

Might also need a reminder of its advantages

Advantages of Corpus Phonetics

automation

Benefits of automation

- Save time in the long run
- Consistency: minimize human error
- Replicability: allow others to repeat the process *identically*
- Easily correct mistakes
- Easily process large amounts of data

Common complaint:

but you don’t get to know your data!!

😑

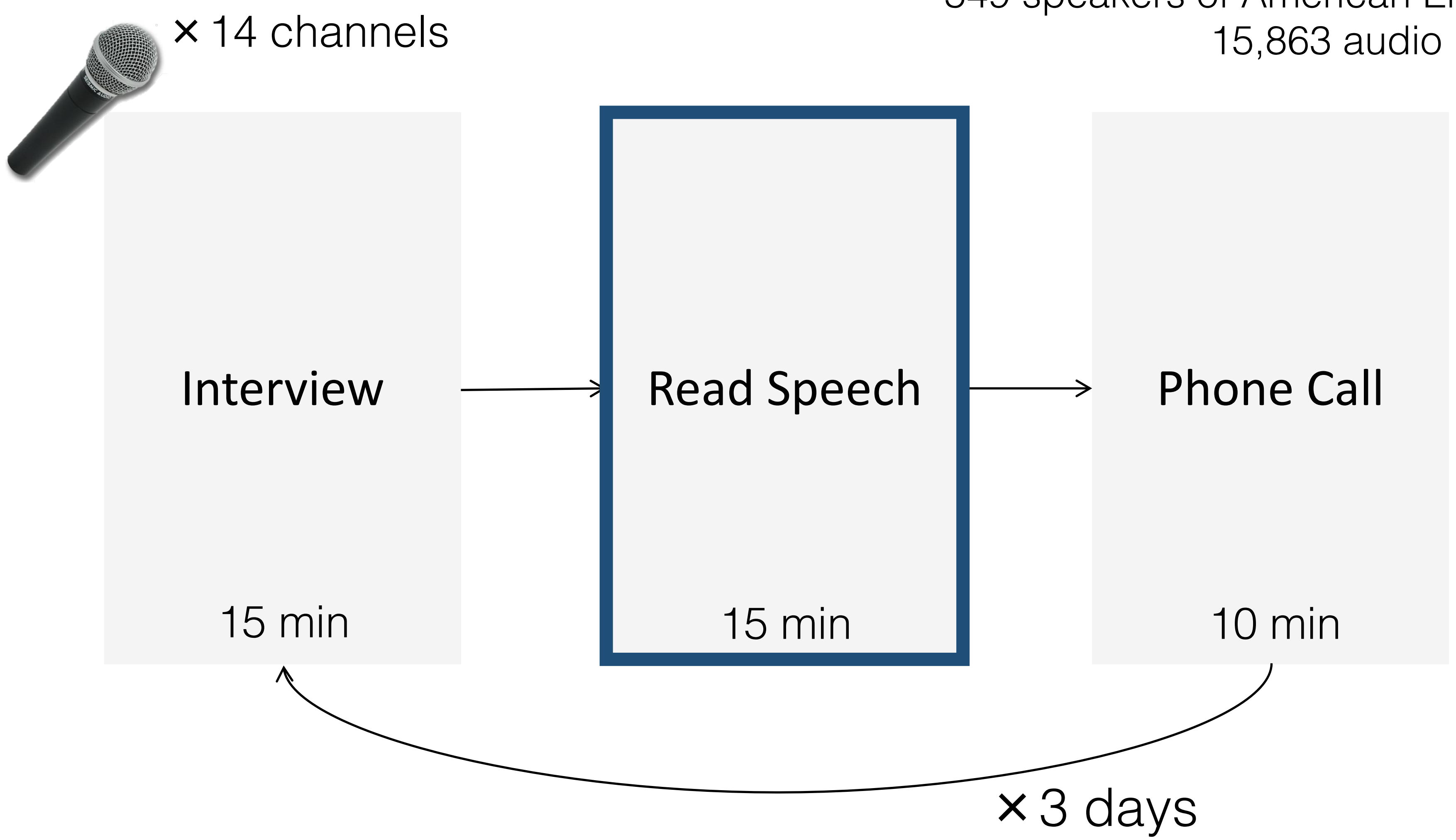
We can still be good scientists and automate

- Automation requires detailed knowledge of what the data looks like
- Automation can also enable better familiarity with the data
- Always audit data (sometimes our coding skills aren’t 💯 or the data is odd, and that’s all ok; also the processing tools aren’t perfect)

Becoming a corpus phonetician

Mixer 6 Corpus

549 speakers of American English
15,863 audio hours



× 14 channels

Interview

15 min

Read Speech

15 min

Phone Call

10 min

× 3 days

Available from the LDC: LDC2013S03

Mixer 6 Corpus: Read Speech

Fixed sentence list with fixed order

Sentences selected from the Switchboard Corpus

Sentence length: 1-17 words (median: 7)

Average 225 sentences per session = ~335,000 read sentences

Unknown number of errors

Unknown locations of errors

Mixer 6 Corpus: Long story short

Mixture of automatic and manual methods to identify location of reading errors and remove

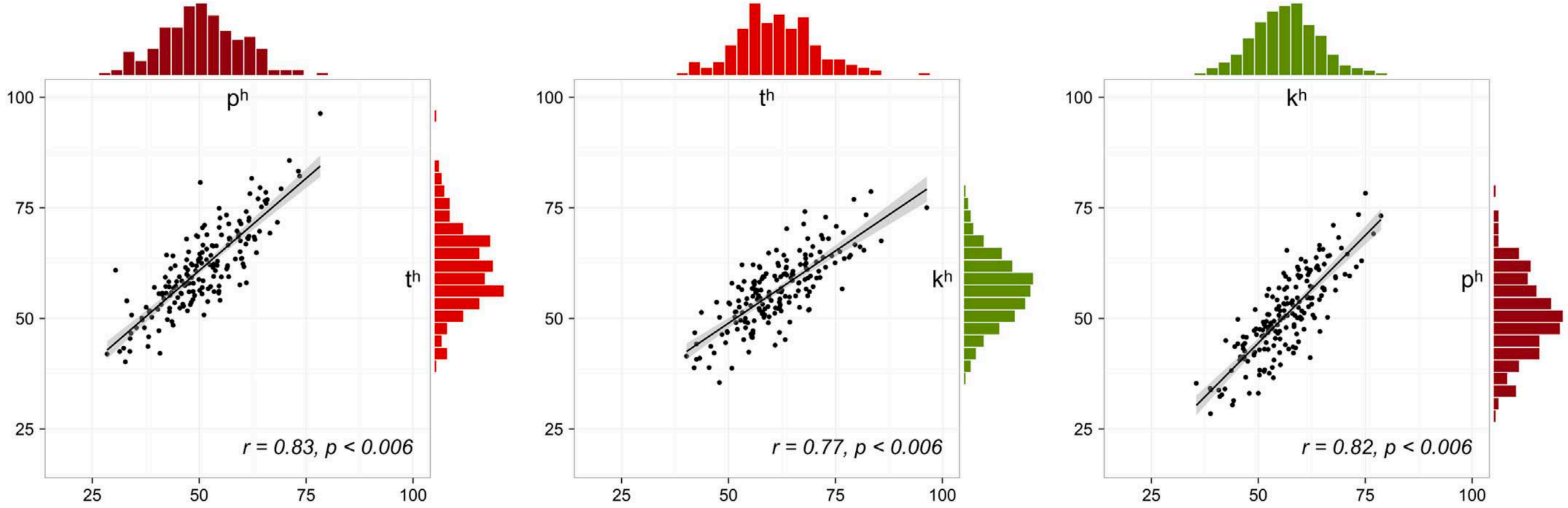
Run forced alignment on “good readings”

Run AutoVOT on word-initial stop consonants

Hand check subset of stop alignments

Mixer 6 Corpus: Long story short

Analyse stop VOT across 180 speakers
88,725 stop consonants
~547 stops per talker



Chodroff & Wilson, 2017

You can do this too!

Workshop goals:

- Become comfortable giving scripting a shot
- Gain basic skillset for doing corpus phonetics
- Implement a pipeline for automatically analysing stops and fricatives

Loose plan

- Command line (session 1, session 2)
- Montreal Forced Aligner intro (session 2)
- Praat Scripting (session 2)
- Montreal Forced Aligner advanced (session 3)

Loose plan

- AutoVOT (session 4)
- Stop and fricative measurement with Praat and R (session 5)
- Overflow, practice, questions, miscellaneous (session 6)

Loose plan

Prepare for a lot of command line use and Praat scripting ☺

Caveats

Workshop will be more applied than theoretical

Translation between Mac and PC

Cygwin

There will be errors: please be patient with me, yourself, and your colleagues → very, very natural part of programming

Installation

Will try to catch some installation errors here, but more likely
we’ll have to spend some time at the beginning of each session

Windows users: Cygwin, does it work? Try vi and vim

Everyone: Montreal Forced Aligner?

Mac users: AutoVOT (it’s a pain of an installation)
Need Xcode and pip, will need to compile program as well