Language, Linguistics, and the Data Explosion:

The Golden Age
of
Speech and Language Studies

Mark Liberman
University of Pennsylvania

http://ling.upenn.edu/~myl

Why the 21st Century is like the 17th Century:

From the perspective of a linguist, today's vast archives of digital text and speech, along with new analysis techniques and inexpensive computation, look like a wonderful new scientific instrument, a modern equivalent of the 17th-century invention of the telescope and microscope.

We can now observe linguistic patterns in space, time, and cultural context, on a scale three to six orders of magnitude greater than in the past, and simultaneously in much greater detail than before.

“Breakfast experiments”

- Our telescope and microscope are
  - Easily available collections of speech and text
  - Computer algorithms for
    - analyzing speech and text
    - collecting, displaying, analyzing statistics
- When we point these new instruments in almost any direction, we see interesting new things
- This is so easy and fast that we can often do an “experiment” on a laptop over breakfast.

These quick looks are not a substitute for serious research.

But they illustrate the power of our new tools,
and allow us to explore interesting new directions quickly.

(For more Breakfast Experiments™ see Language Log)

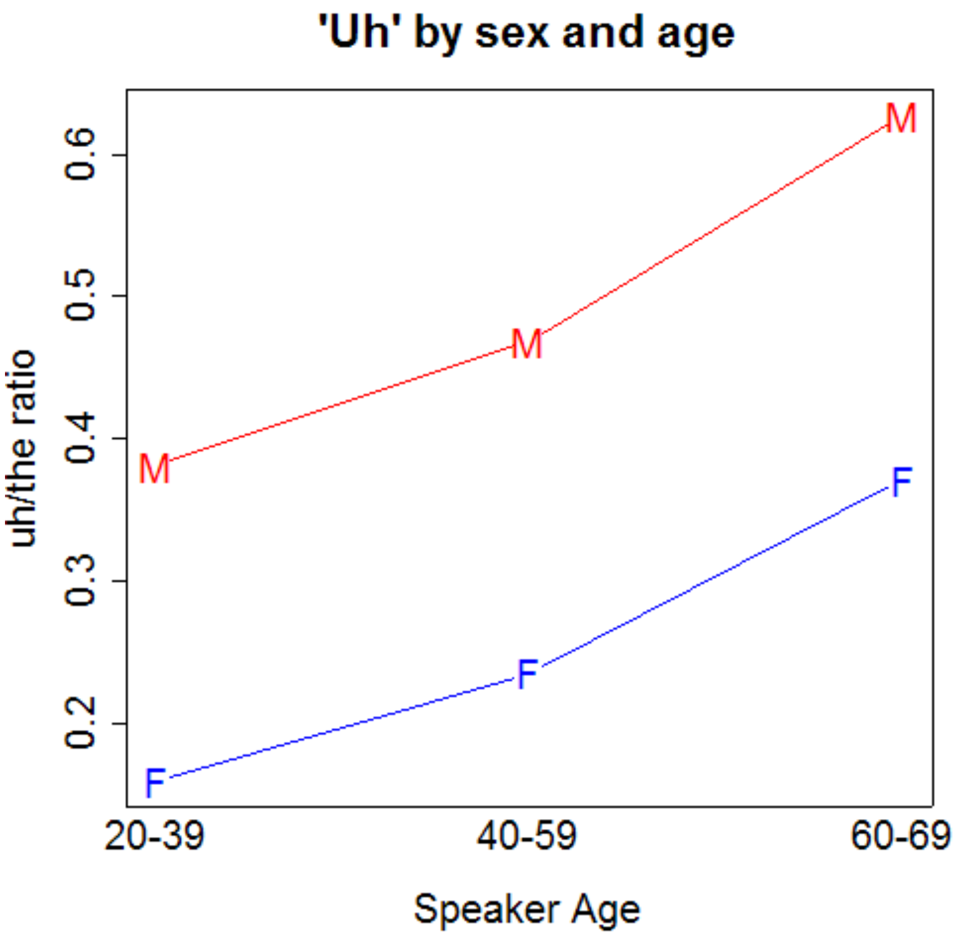
A Sample Breakfast Experiment

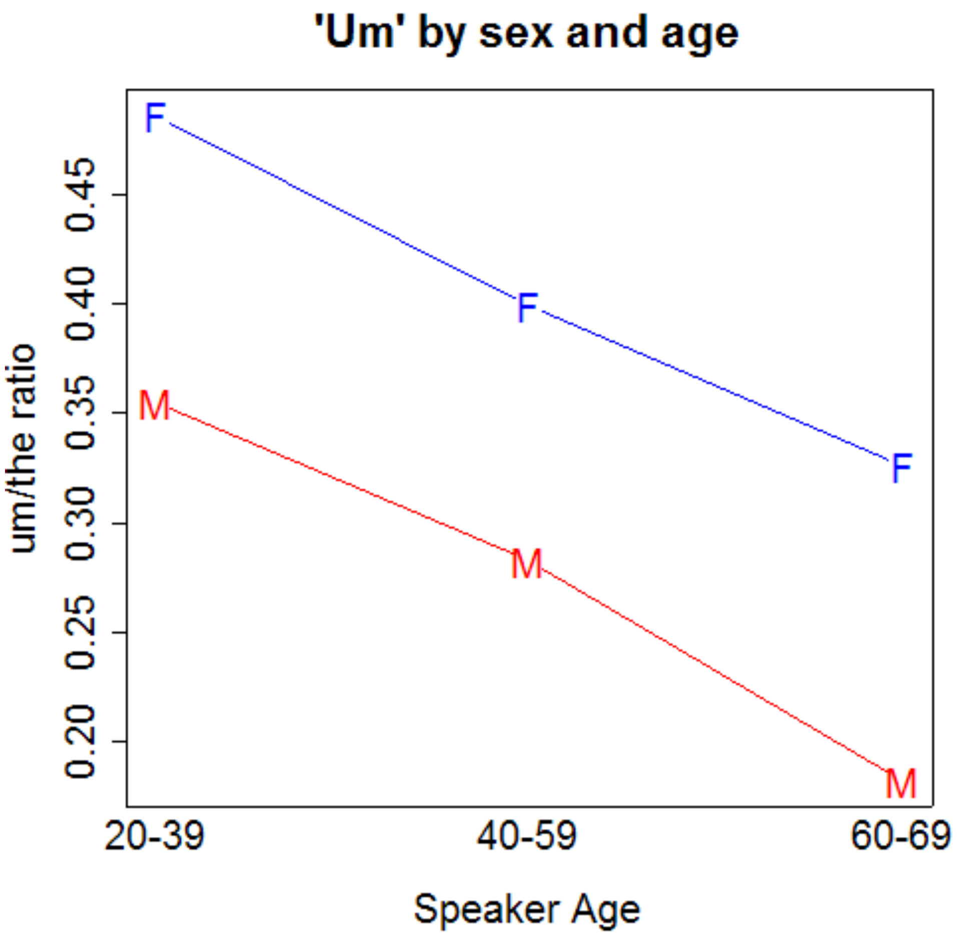
- How does disfluency vary with sex and age?
- Method: count ums and uhs

  in transcripts of U.S. English conversations

  by demographic categories of speakers
- Data: 14,500 recorded and transcribed telephone conversations

  (2,500 hours of audio; 28 million words)
- Result: systematic but unexpected interaction


'Uh' by sex and age
uh/the ratio
0.2
0.3
0.4
0.5
0.6
M
M
M
F
F
F
20-39
40-59
60-69
Speaker Age


'Um' by sex and age
um/the ratio
0.20
0.25
0.30
0.35
0.40
0.45
F
F
F
M
M
M
20-39
40-59
60-69
Speaker Age

Time to get the numbers: 60 seconds

Time to make the graphs: 5 minutes

Time to write the blog post: 45 minutes

Why the 21st Century is Also Like the 16th Century:

Research datasets are no longer the exclusive preserve of the scientific hierarchy –

Any bright undergraduate with an internet connection can access and interpret the primary data for herself.

At least, she can do so sometimes…

This is still contested by the hierarchy,

but the reformation is winning!

Why this is transformative:

When research datasets are available, there's more research,
because barriers to entry are lowered.

When research datasets are shared, the research is better,
because results can be replicated,
and algorithms and theories can be compared.
In addition, shared datasets are much bigger and more expensive
than any individual researcher's time and money would permit.

And when datasets are associated with well-defined research questions,
the whole field gets better,
because the people who work on a "common task"
form a community of practice
within which ideas and tools circulate rapidly.

It’s happened before:

This is not a new set of ideas.

Europe made an analogous set of discoveries in the 16th century, when the printing press transformed European society.

Literacy, education, and scholarship spread much more widely, and improved in quality as well as quantity along the way.

We might call this process the "Data Reformation",
since it emphasizes the spread of unmediated access
to the primary material needed to discover truth.

More familiar names for the trend are the "Open Data"
and "Reproducible Research" movements.

Under whatever name,
this trend is making increasing amounts of digital data –
including speech and language data –
accessible to increasingly many researchers worldwide.

There are issues with protecting privacy and property rights,
and with paying for creation, duration, and distribution.
But solutions exist or can be devised.

I came here from IEEE ICASSP 2014
(International Conference on Acoustics, Speech, and Signal Processing)
in Italy. Here's a typical note from a paper presented there:

> 6. REPRODUCIBLE RESEARCH
> This research benefits from the efforts of other researchers
> to share their code [5] and datasets [21].
> The open availability of these resources is commendable,
> allowing other researchers to easily and accurately compare methods.
> The code used in the experiments described in this paper is available
> at http://code.soundsoftware.ac.uk/projects/gs bnmf/.

More than 90% of the papers I attended there used published digital data,
and many also shared their software.

The engineers are leading the way in this revolution,
but linguistic science and scholarship are being transformed as well.

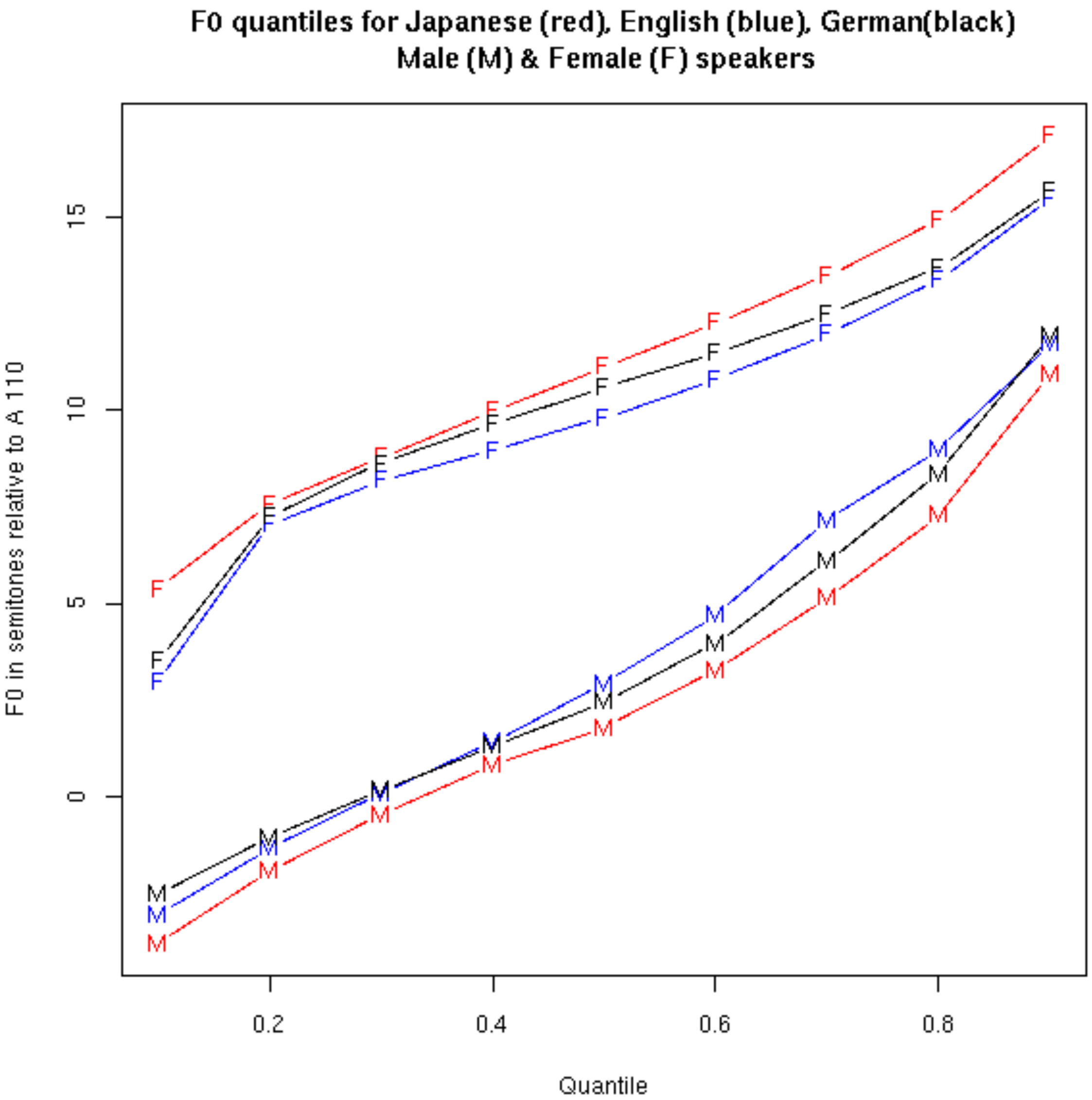
Thank you!

Five More Breakfast Experiments

- Do Japanese speakers show more gender polarization in pitch than American speakers?
- Do American women talk more (and faster) than men?
- How does word duration vary with phrase position?
- How does local speaking rate vary in the course of a conversation?
- “you know”/”I mean” ratio over the lifespan

Breakfast Experiment #1

- Gender polarization in conversational speech
- Question: are Japanese men and women more polarized (more different) in pitch than Americans or Europeans?
- Method:
  - Pitch-track published telephone conversations
  - LDC "Call Home" publications for Japanese, U.S. English, German
    - Collected 1995-1996 , published 1996-1997
    - about 100 conversations per language
  - Compare quantiles of pooled values (about 2 million numbers per sex/culture combination)
- Answer: yes, apparently so.



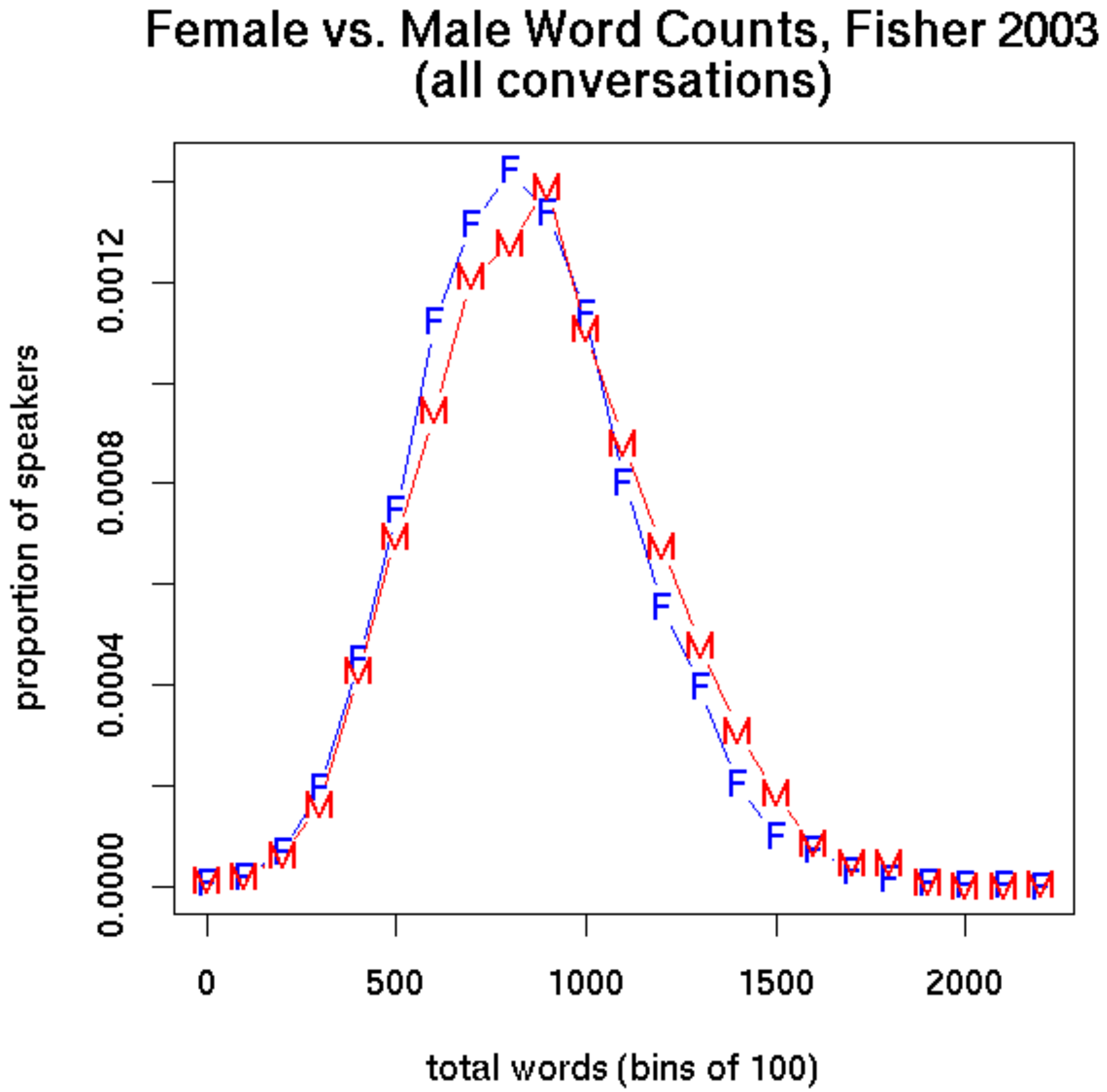
Data from CallHome M/F conversations; about 1M F0 values per category.

As usual, more questions:

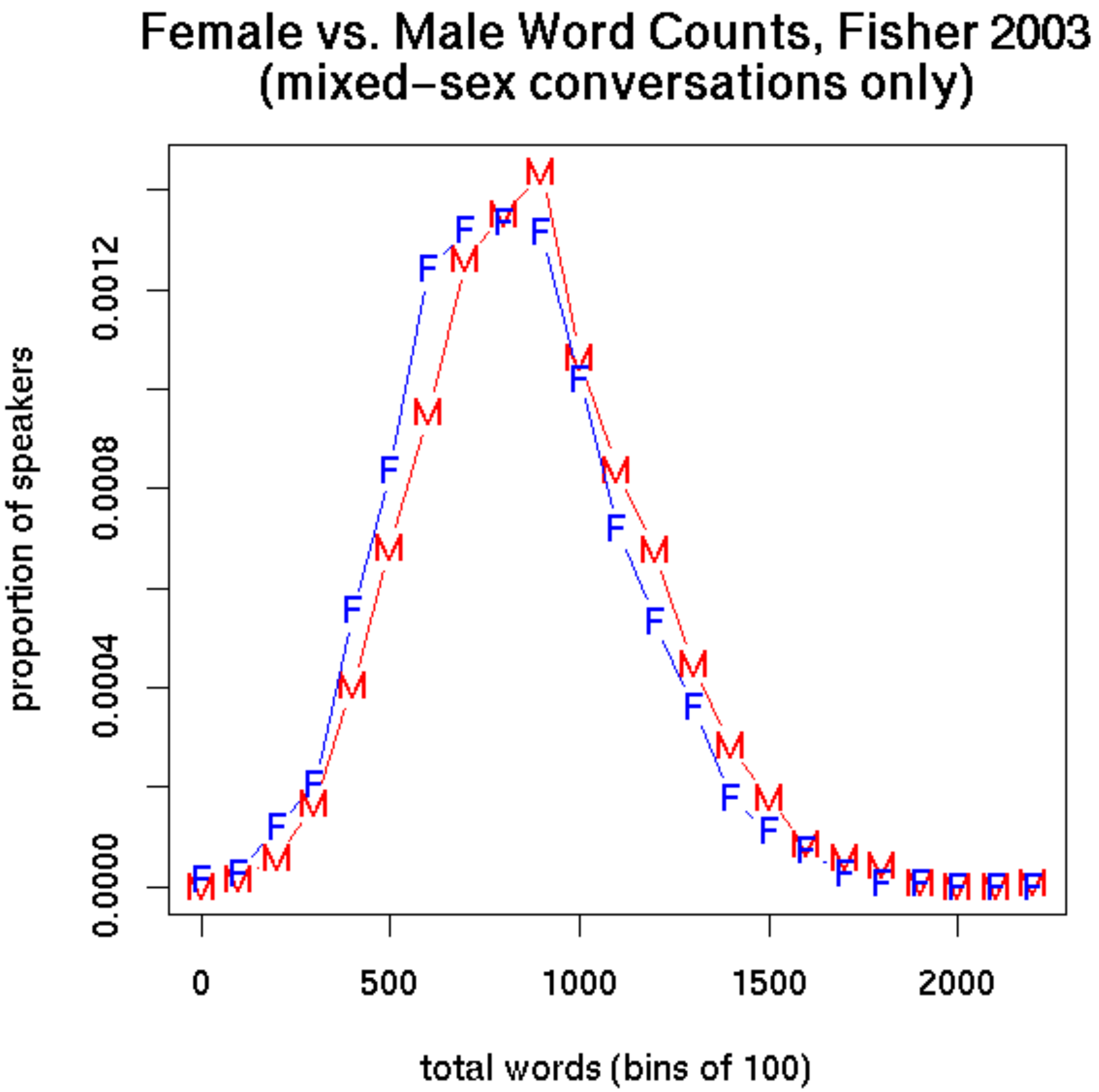
- Other cultures and languages
- Effects of speaker's age
- Effects of relationship between speakers, nature of discussion
- Formal vs. conversational speech
- Effects of social class, region

Experiment #2a

- Sex differences in conversational word counts
- Question: Do women talk more than men?
- Method: Count words in "Fisher" transcripts
  - Conversational telephone speech
    - Collected by LDC in 2003
    - 5,850 ten-minute conversations
      - 2,368 between two women
      - 1,910 one woman, one man
      - 1,572 between two men
- Answer: No.

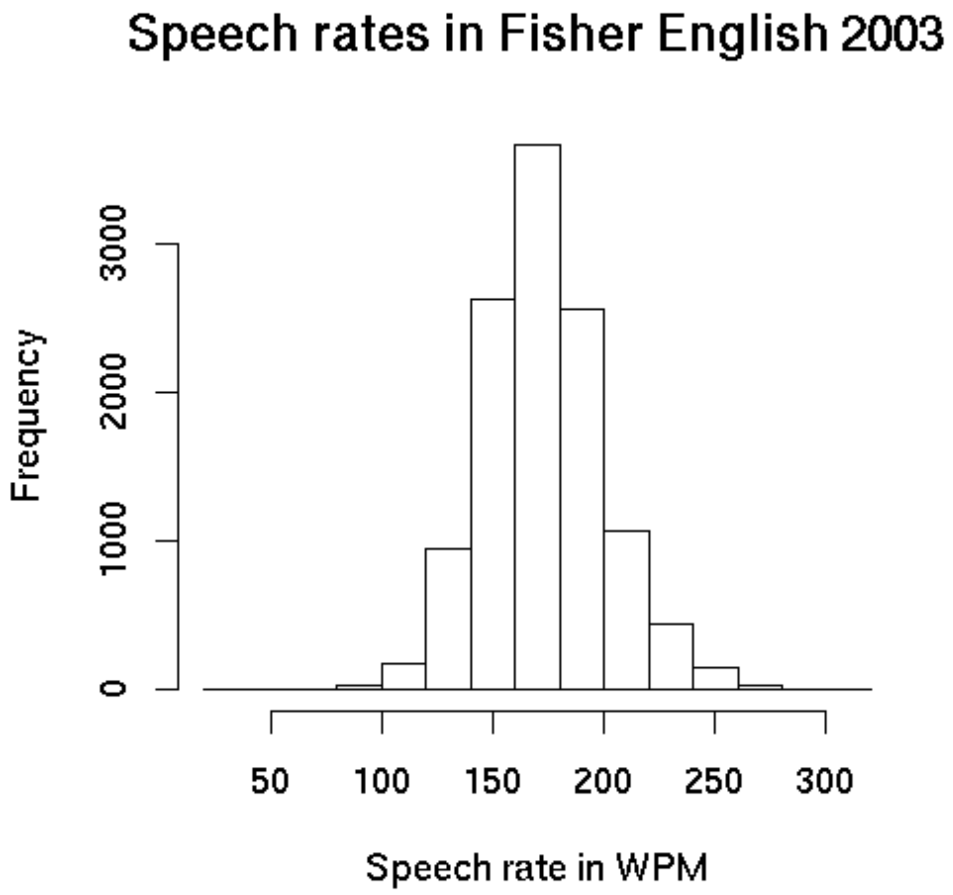


Female vs. Male Word Counts, Fisher 2003
(all conversations)


Female vs. Male Word Counts, Fisher 2003
(mixed−sex conversations only)
proportion of speakers
0.0000
0.0004
0.0008
0.0012
0
500
1000
1500
2000
total words (bins of 100)
F
M

Experiment #2b

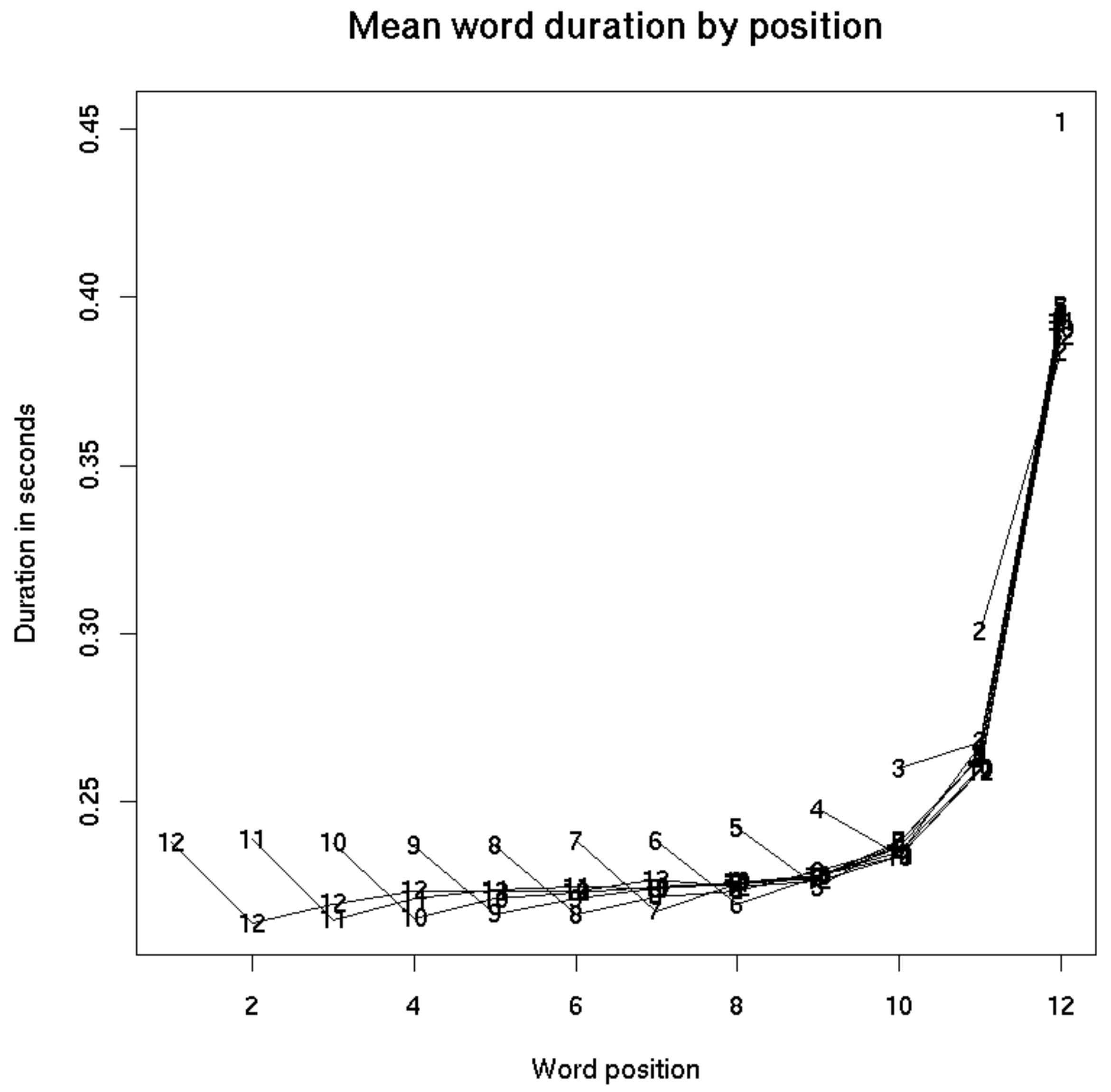
- Sex differences in conversational speaking rates
- Question: Do women talk faster than men?
- Method: Words and speaking times in Fisher 2003
- Answer: No.



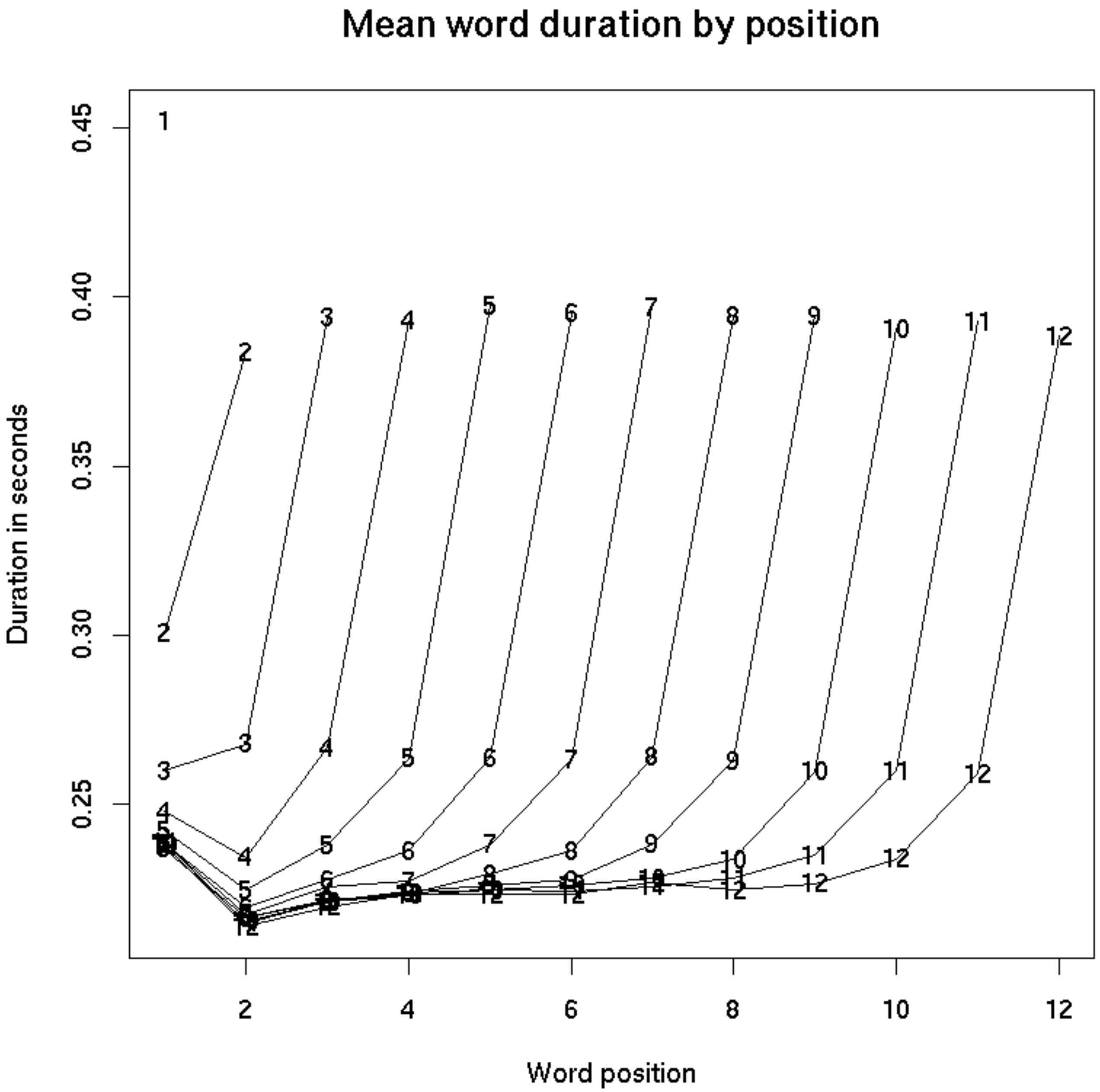
(11,700 conversational sides; mean speaking rate=173 wpm, sd=27)
(Male mean 174.3, female 172.6: difference 1.7, effect size d=0.06)

Experiment #3

- Phrasal modulation of speaking rate
  - "final lengthening" a well-established effect
  - first observed by Abbé J.-P. Rousselot ~1870
  - other phrase-position effects are less clear
- What is a "phrase"?
  - A syntactic unit?
  - A unit of information structure?
  - A unit of speech production?
- Method: word duration by position in "pause group" (stretch of speech without internal silence >100 msec)
- Data: Switchboard corpus
- Result: Amazingly regular (average) pattern

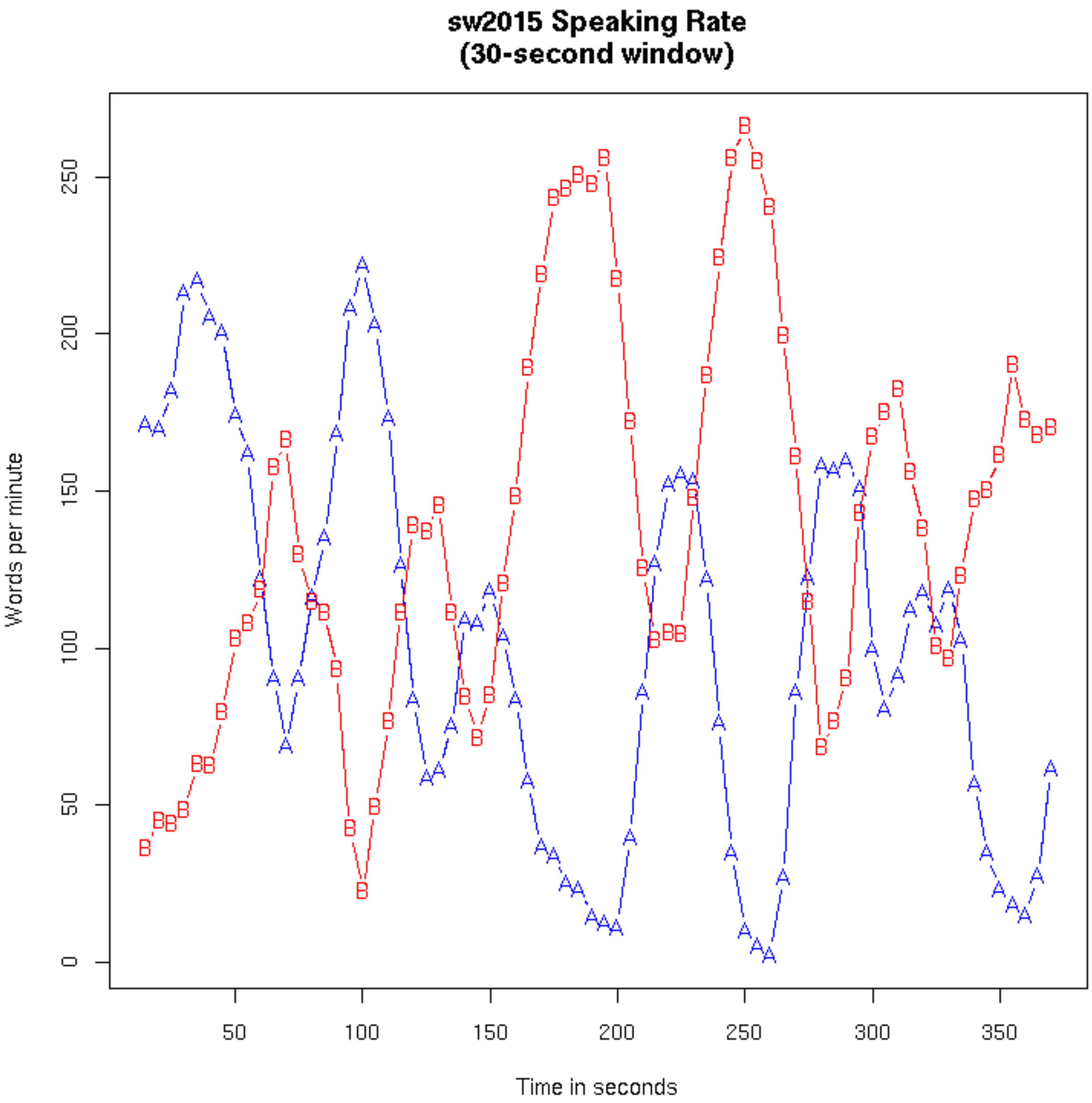


Data from Switchboard; phrases defined by silent pauses
(Yuan, Liberman & Cieri, ICSLP 2006)


Mean word duration by position
Duration in seconds
0.45
0.40
0.35
0.30
0.25
Word position
2
4
6
8
10
12

Experiment #4

- How does speaking rate
  reflect the ebb and flow of a conversation?
- Method: word- or syllable-count
  in moving window
  over time-aligned transcripts
- Result: suggestive pictures


sw2015 Speaking Rate
(30-second window)
Words per minute
0
50
100
150
200
250
Time in seconds
50
100
150
200
250
300
350
A
B

Experiment #5

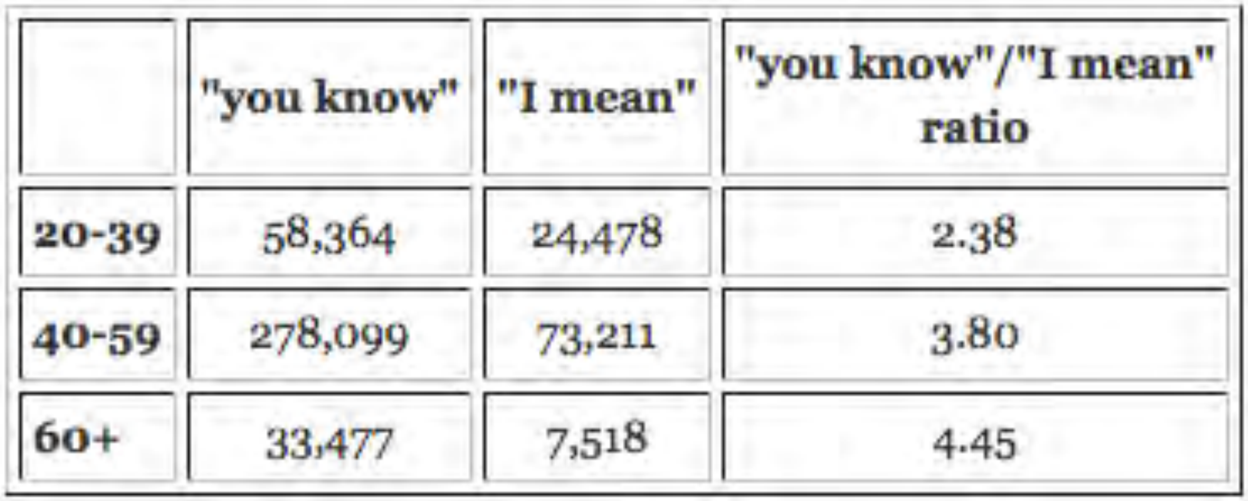
The News Editor at Psychology Today wrote to me:

> Sometimes I wonder if there are underlying personality differences between people who punctuate (litter?) their speech with "you know" versus those who use "I mean" more frequently. Any hunch on that?

I didn’t have any hunches, and there didn’t seem to be anything relevant in the literature. But I did have access to an indexed copy of the the 14,137 conversations (26,151,602 words) in the LDC’s English-language conversational speech corpora.

(…and so do you!)

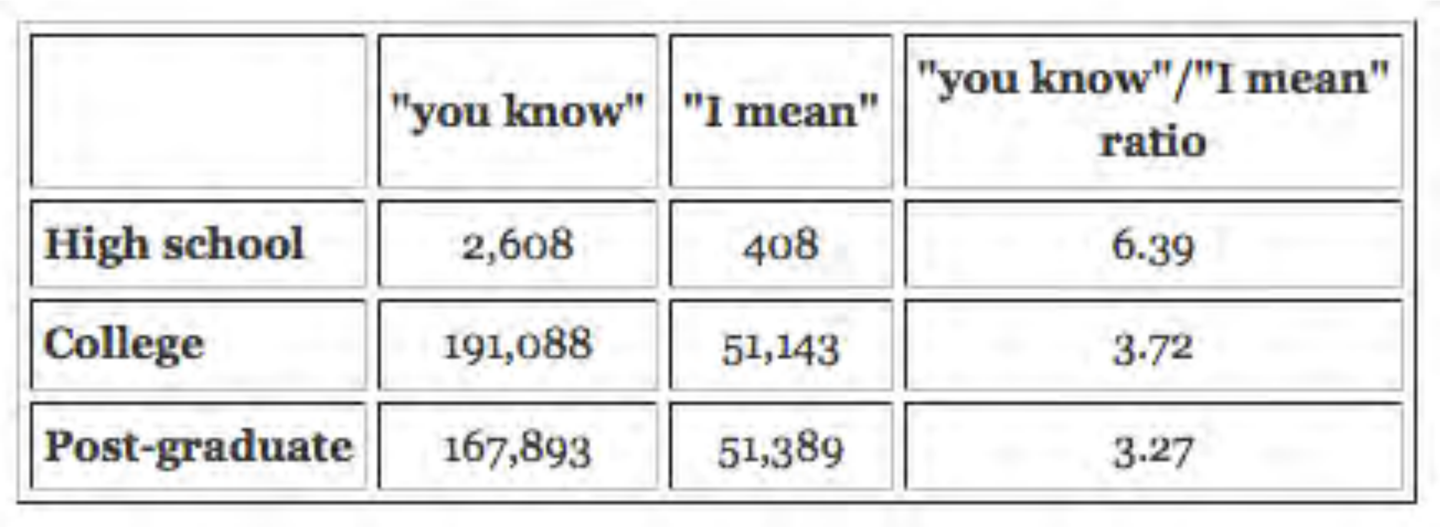
And there’s demographic data for (almost) all speakers.
So I checked:



	"you know"	"I mean"	"you know"/"I mean" ratio
20-39	58,364	24,478	2.38
40-59	278,099	73,211	3.80
60+	33,477	7,518	4.45

Elapsed time: 6 queries + 3 ratio calculations = 5 minutes

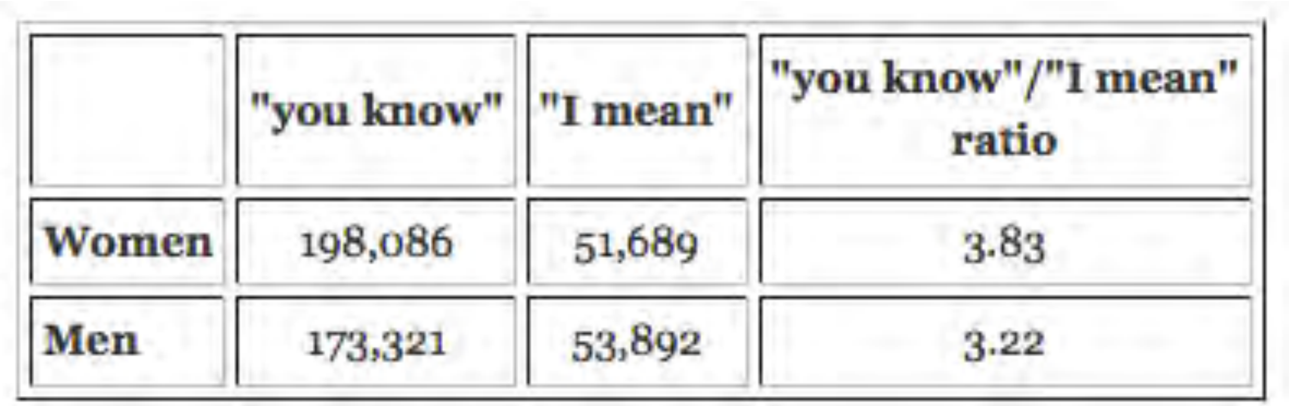
What about the effect of years of education?



	"you know"	"I mean"	"you know"/"I mean" ratio
High school	2,608	408	6.39
College	191,088	51,143	3.72
Post-graduate	167,893	51,389	3.27

(Caveat:
High-school-only group was small,
and perhaps mainly older…)

Sex differences?



	"you know"	"I mean"	"you know"/"I mean" ratio
Women	198,086	51,689	3.83
Men	173,321	53,892	3.22

Elapsed time:

15 minutes for queries, 45 minutes to write it up

("I mean, you know", Language Log, 8/19/2007)

My conclusion

Maybe greater use of "I mean" means greater involvement with self as opposed to others, and that age makes people less self-involved, but education makes them more self-involved, and men are somewhat more self-involved than women.

But this is even more tenuous than such explanations generally are, since the demographic variables in this collection of conversations are not orthogonal.

So you'd want to do some sort of hierarchical regression, and it would take a day or two to get the data and run it.

But still **. . .**

Serious speech science

- Transcribed speech
  is available in very large quantities
- By applying
  - forced alignment
  - pronunciation modeling
  - automated measurements
  - multilevel regression

  we see a new universe of speech data,
  on a scale 4-5 orders of magnitude
  greater than the laboratory recordings
  of the past.

- And interesting patterns are everywhere!

Interdisciplinary opportunities

- These techniques will have rich applications in other fields
  - Clinical diagnosis and evaluation
  - Educational assessment
  - Social science survey methods
  - Studies of performance style
  - . . . and so on . . .
- Wherever speech and language are relevant!

Even in classical scholarship!

The early years of the twenty-first century have seen a heroic age for intellectual life. Ideas have poured across the world and new minds have joined the professionalized academics and authors in grappling with the heritage of humanity. […]

No field of study is poised to benefit more than those of us who study the ancient Greco-Roman world and especially the texts in Greek and Latin to which philologists for more than two thousand years have dedicated their lives. […]

The terms eWissenschaft and ePhilology, like their counterparts eScience and eResearch, point towards those elements that distinguish the practices of intellectual life in this emergent digital environment from print-based practices. Terms such as eWissenschaft and ePhilology do not define those differences but assert that those differences are qualitative. We cannot simply extrapolate from past practice to anticipate the future.

-- Gregory Crane et al., “Cyberinfrastructure for Classical Philology”,
Digital Humanities Quarterly, Winter 2009

An historic opportunity:

- Take an interesting problem, and add
  - a little linguistics and phonetics
  - a little psychology
  - a little signal processing
  - a little statistics and machine learning
  - a little computer science
  - your curiosity and initiative
- And the future is yours!