The Future of Computational Linguistics: or, What Would Antonio Zampolli Do?

Mark Liberman

University of Pennsylvania

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

Antonio Zampolli’s history

- Statistical lexicography (Thomas Aquinas)
- 1969: Centro Nazionale Universitario di Calculo Elettronico at the University of Pisa
- 1970s: “*International Summer Schools in Computational and Mathematical Linguistics*”
- 1980s and onward: many multi-site European and international projects: EUROTRA etc. etc.,

which established “Language Engineering” in Europe

Legacy of the Pisa meetings

They created an enduring community.
Most older people here today
participated in them.
Most younger people here today
were taught by someone who participated,
or were taught by someone who was
taught by someone who did.

Antonio’s Outlook

Pessimist: the glass is half empty.

Optimist: the glass is half full.

Antonio:

- We have a great opportunity!

  our glass is empty

  . . . and Brussels has a bottle!

He was a great intellectual entrepreneur.

So What Would Antonio Say Now?

"Let us

re-invent

the sciences

of speech and language!"

Support for this view from the U.S. National Academy of Sciences:

We see that the computer has opened up to linguists a host of challenges, partial insights, and potentialities. We believe these can be aptly compared with the challenges, problems, and insights of particle physics. Certainly, language is second to no phenomenon in importance. And the tools of computational linguistics are considerably less costly than the multibillion-volt accelerators of particle physics. The new linguistics presents an attractive as well as an extremely important challenge.

There is every reason to believe that facing up to this challenge will ultimately lead to important contributions in many fields.

Report by the Automatic Language Processing Advisory Committee,
National Academy of Sciences

Two wrinkles

(1) ALPAC ’s main recommendation
was to de-fund Machine Translation research.

(2) And, the ALPAC report came out in **1966** (!)

so 44 years later,
where’s the QCD of computational linguistics?

The plan vs. the reality

- ALPAC ‘s idea:
  1. computers → new language science
  2. language science → language engineering
- What actually happened:
  1. computers → new language engineering
- Today’s opportunity?
  2. engineering → new language science (?)

What went wrong after 1966?

- 1970-era Computers were not enough:
  we also needed
  - adequate accessible digital data
  - tools for large-scale automated analysis
  - applicable research paradigms

- Now we have these.
- (at least, two out of three…)

Hypothesis: 2010 is like 1610

- We’ve invented
  the linguistic telescope and microscope:
  - Inexpensive networked computation
  - Effective and flexible analysis algorithms
  - A growing universe of digital text and speech.
- We can observe linguistic patterns
  - in space, time, and cultural context
  - on a scale 3-6 orders of magnitude greater than before
  - and also in much greater detail.

Now ALPAC’s prediction may come true:

Research that “can be aptly compared with the challenges, problems, and insights of particle physics.”

Of course, that’s what they all say . . .

Progress in any science depends on a combination of improved observation, measurement, and techniques. The cheap computing of the past two decades means there has been a tremendous increase in the availability of economic data and huge strides in econometric techniques. As a result, economics stands at the verge of a golden age of discovery.

-Diane Coyle, “Economics on the Verge of a Golden Age”,
The Chronicle of Higher Education, March 12, 2010

… but maybe it’s true!

- “eScience” is developing in every area:
  - = computationally intensive science
    using immense data sets
    in highly distributed network environments.
- The sciences of speech and language
  are uniquely well positioned
  to use these techniques --
- And also to offer new eScience methods
  to other disciplines.

Interesting patterns are everywhere

- Given a well-organized body of linguistic data,

  many questions
  can be asked and answered easily,

  - with answers that are often unexpected,
    raising new questions of fact and interpretation,
    and opportunities for modeling and explanation.

- Yogi Berra:

  "Sometimes
  you can observe a lot
  just by watching."

A rapid tour of simple examples:

- 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 declination slope vary with phrase length?
- How does local speaking rate vary in the course of a conversation?
- How does disfluency vary with sex and age?

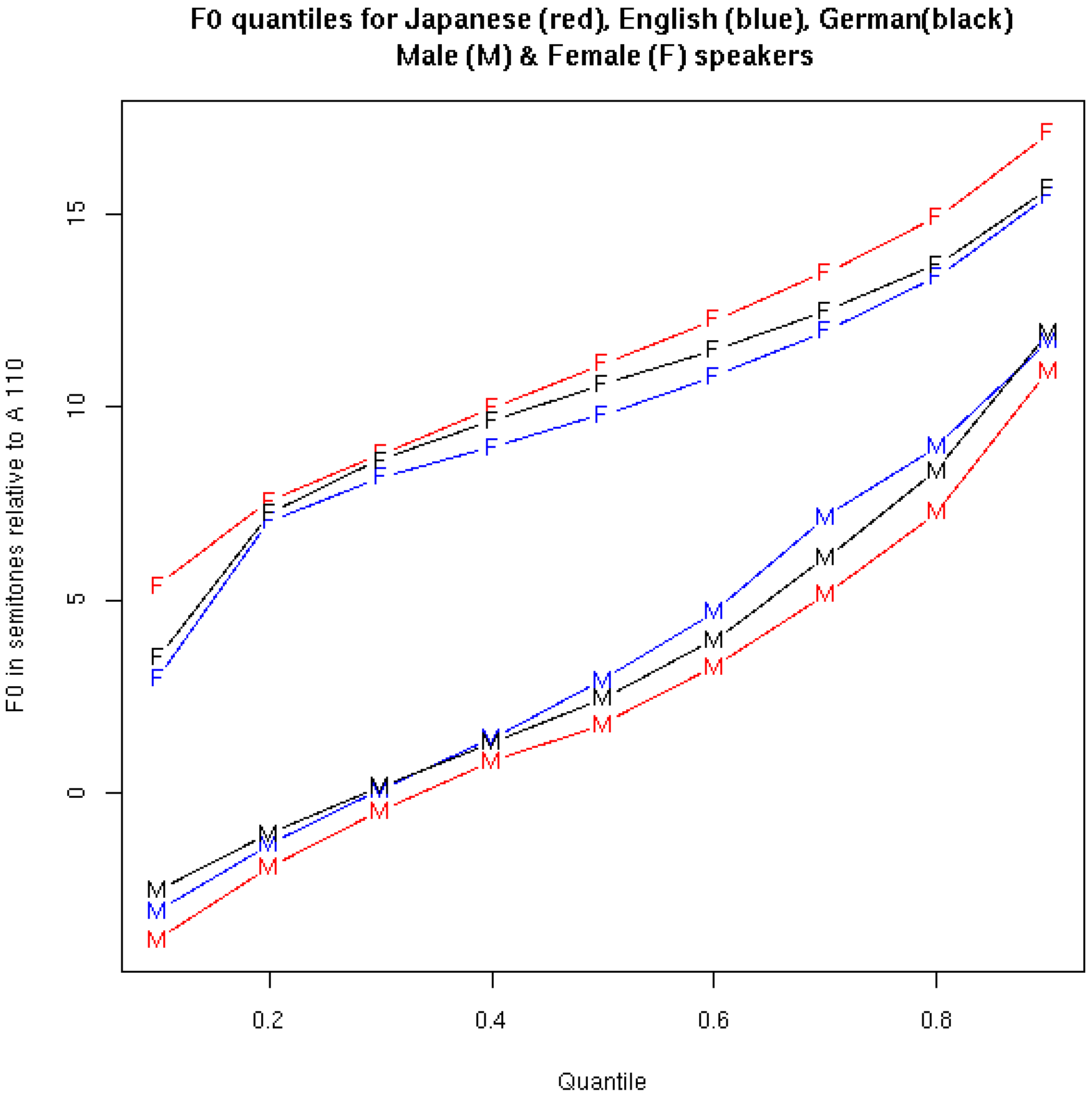
These are illustrative examples
of questions that can be asked and answered in a few minutes
with modern techniques and resources.

Some of them come from larger studies,
with collaborators including Jiahong Yuan among others.

Rather than settling the matter,
each example suggests new questions to investigate.

The point here
is simply that interesting patterns are everywhere you look,
and that large-scale looking
has is now becoming increasingly easy.

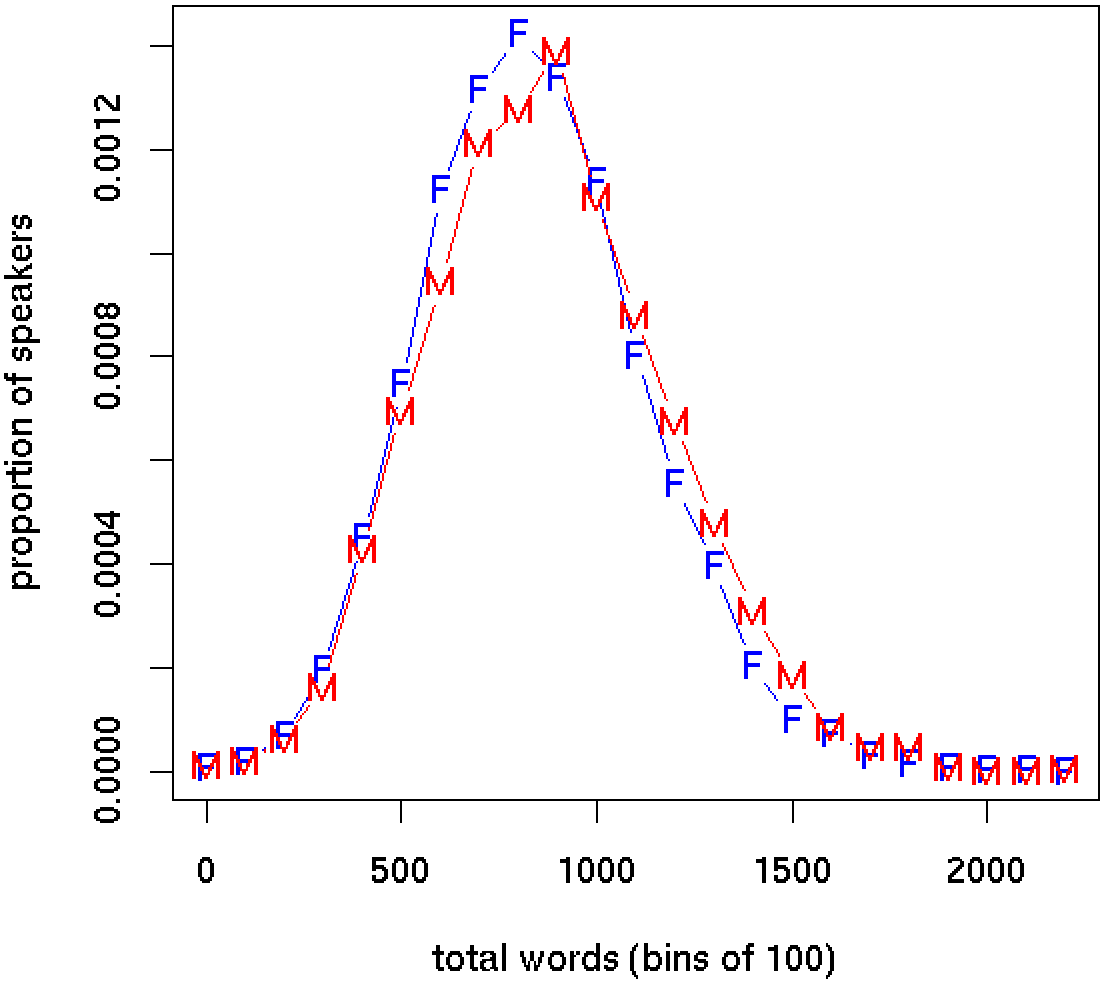
1. Cultural differences in gender polarization of pitch range

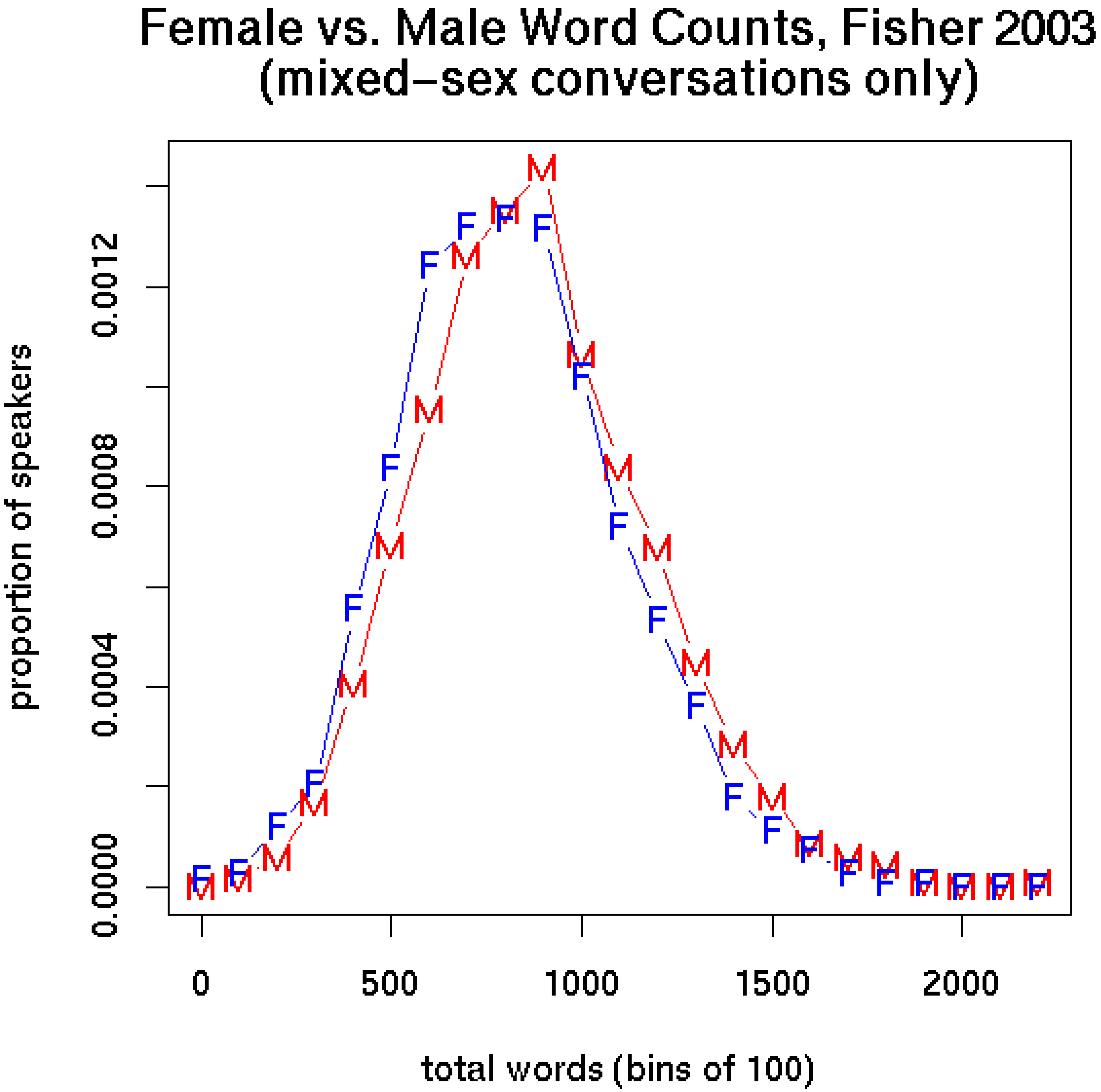


Data from CallHome M/F conversations; about 1M F0 values per category.

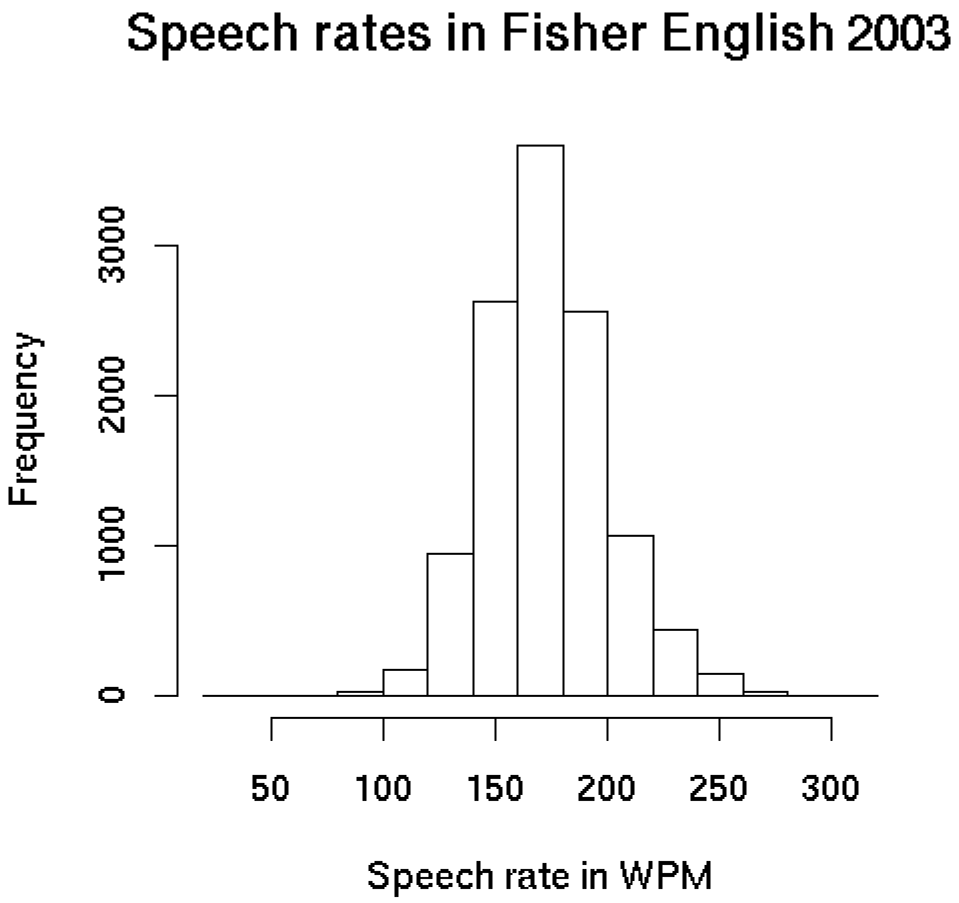
2(a). Sex differences in conversational word counts?

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



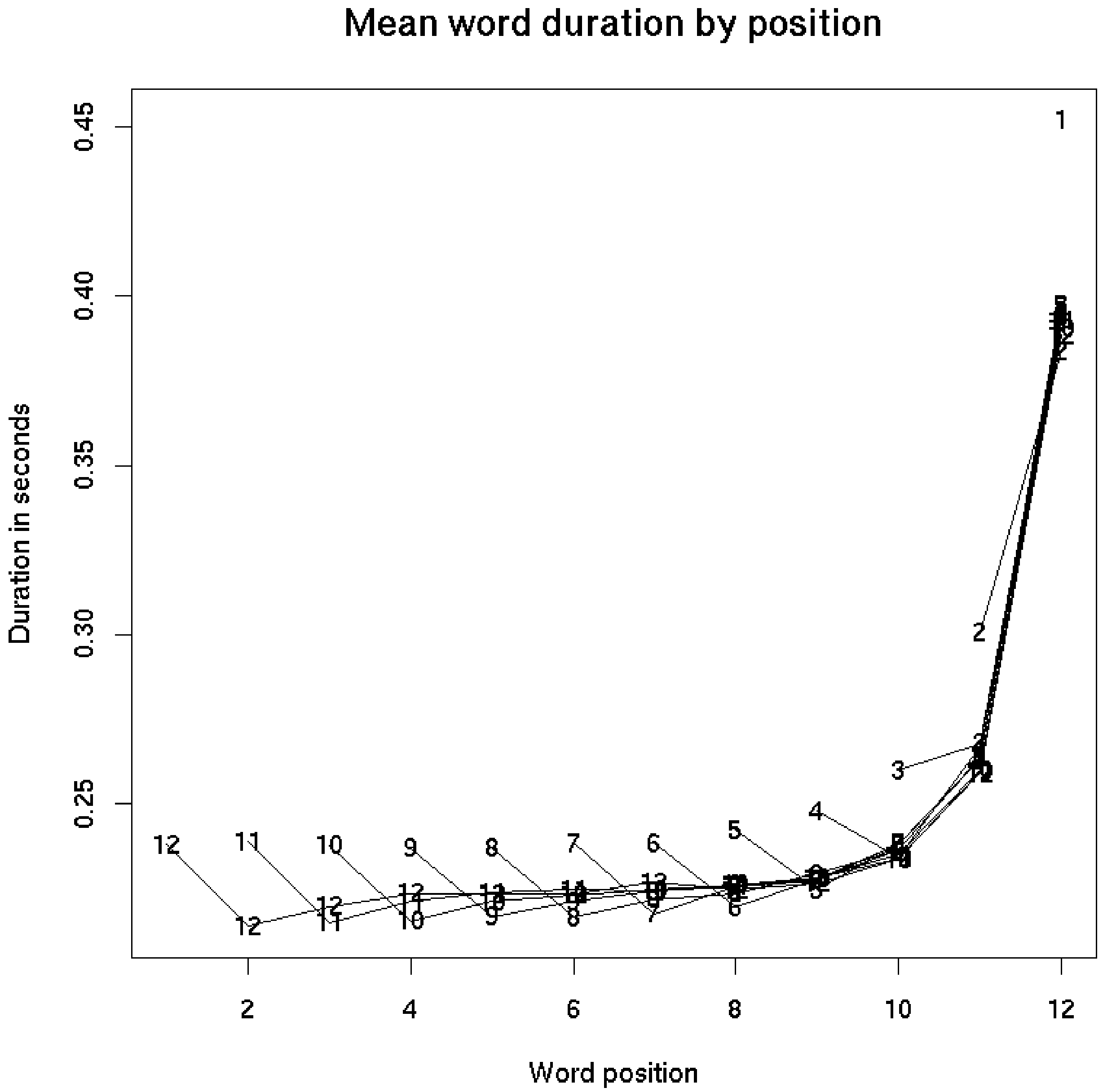

Female vs. Male Word Counts, Fisher 2003
(mixed-sex conversations only)
proportion of speakers
0.0000
0.0004
0.0008
0.0012
0
500
1000
1500
2000
total words (bins of 100)
F
M

2(b). Sex differences in speaking rate?

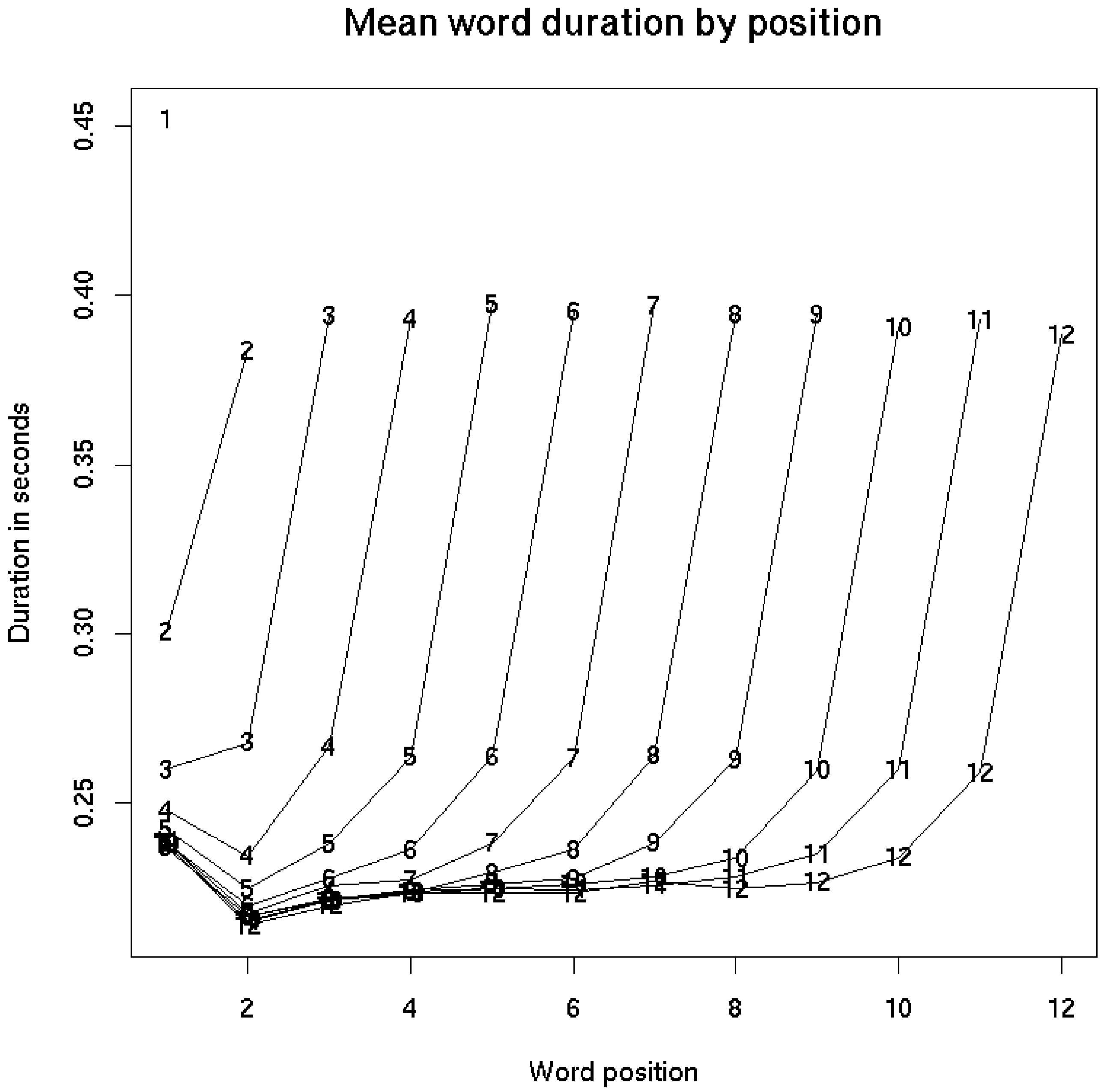


(11,700 conversational sides; mean=173, sd=27)
(Male mean 174.3, female 172.6: difference 1.7, effect size d=0.06)

3. The shape of speech rate in a spoken phrase

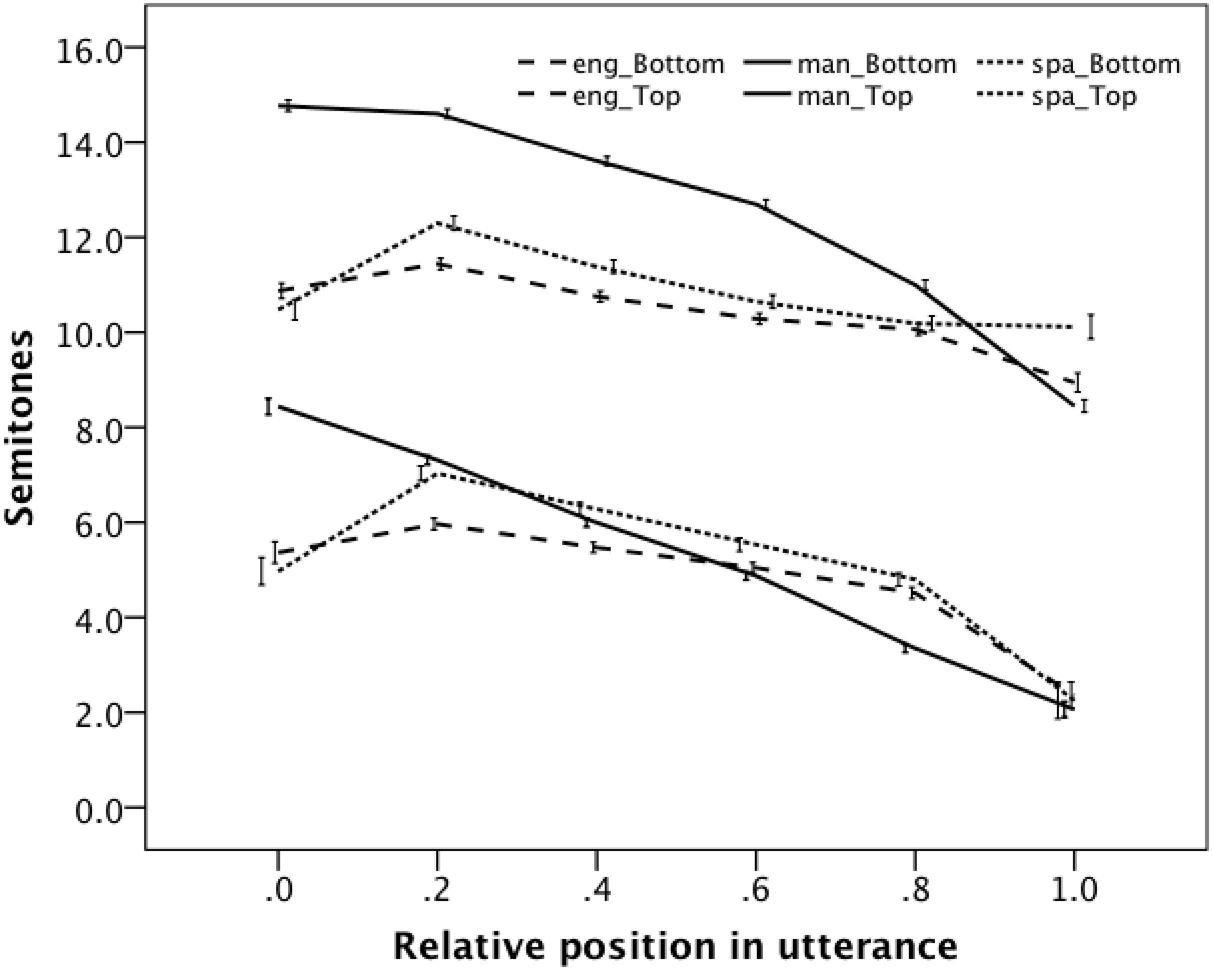


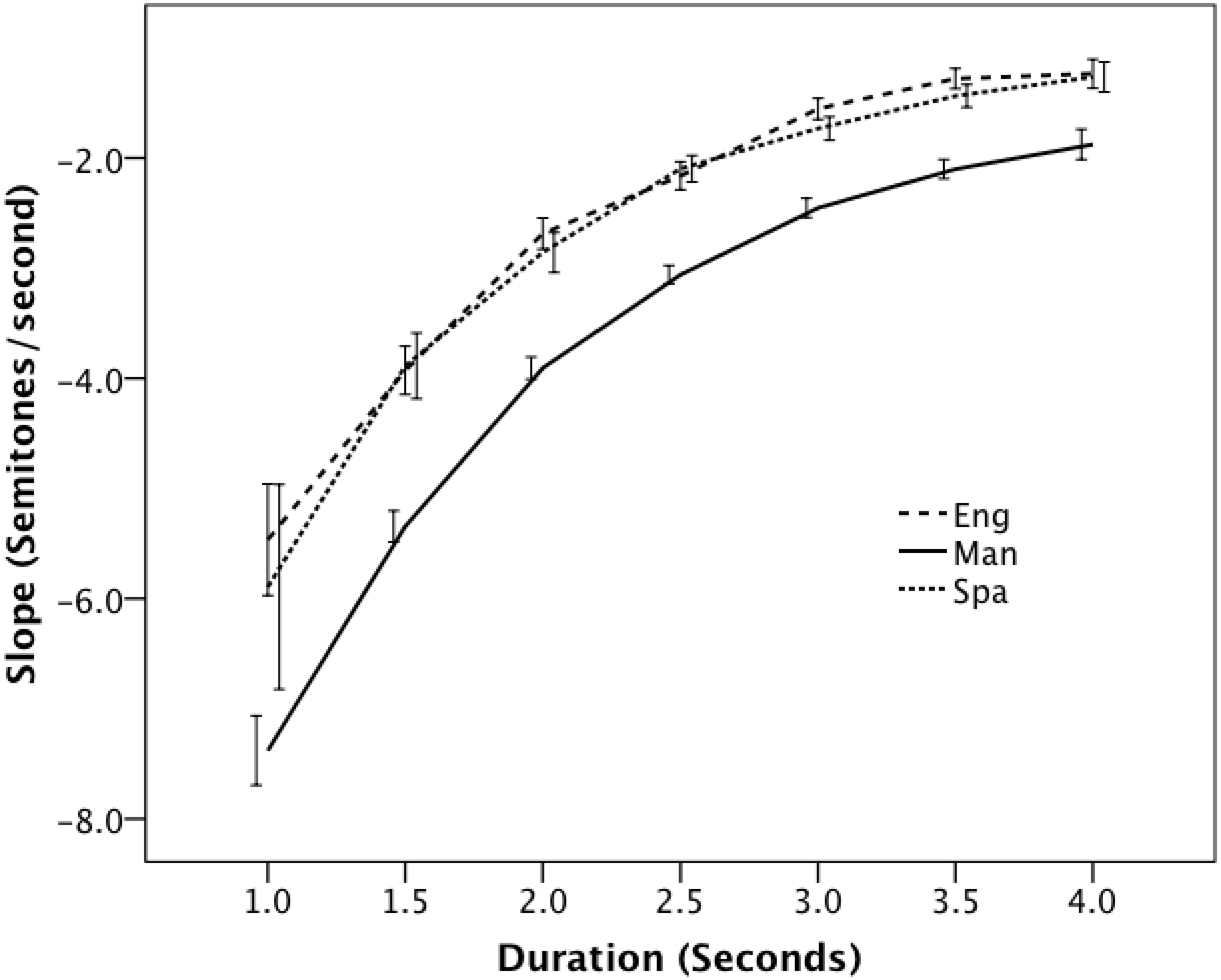
Data from Switchboard; phrases defined by silent pauses
(Yuan, Liberman & Cieri, ICSLP 2006)

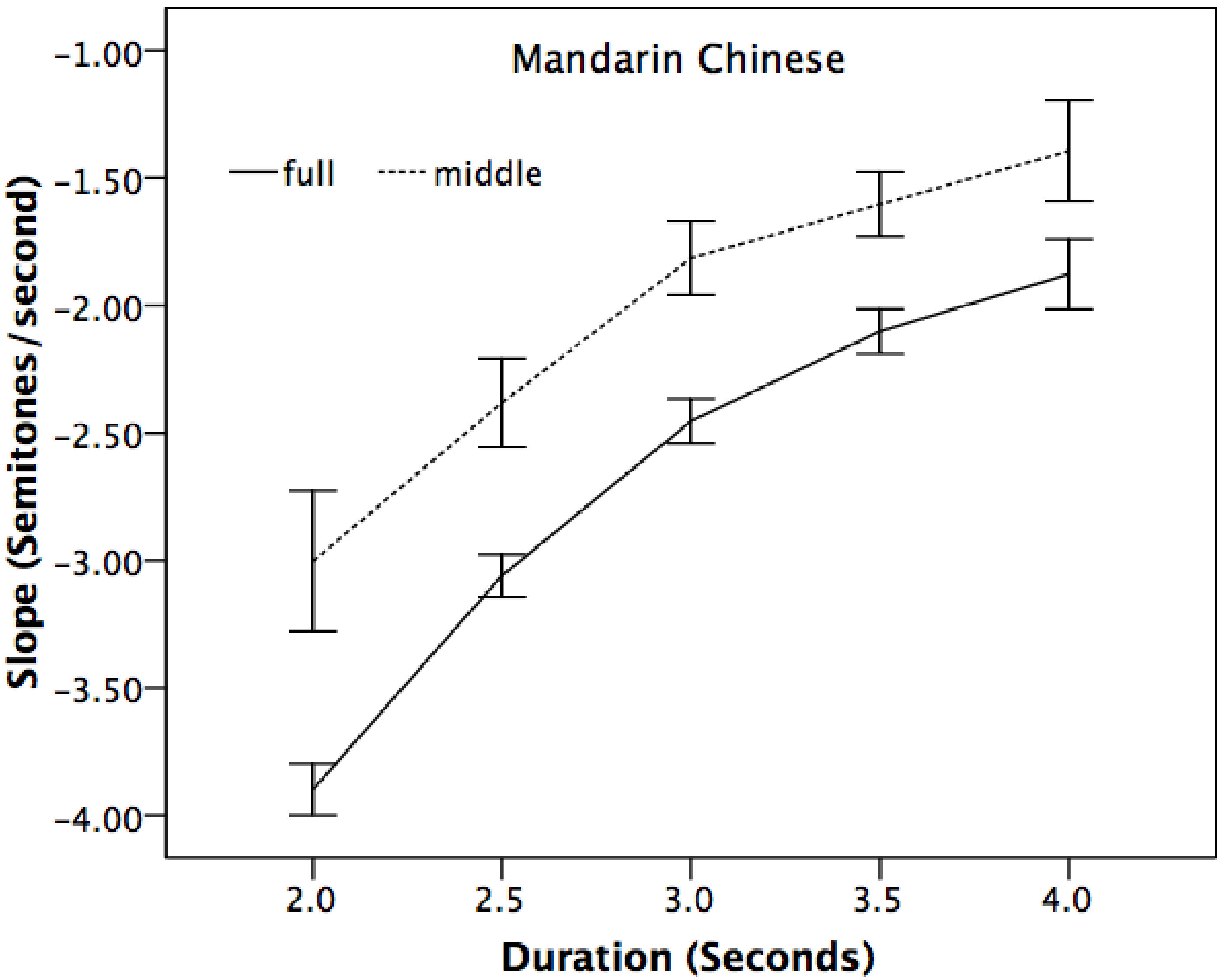


Mean word duration by position

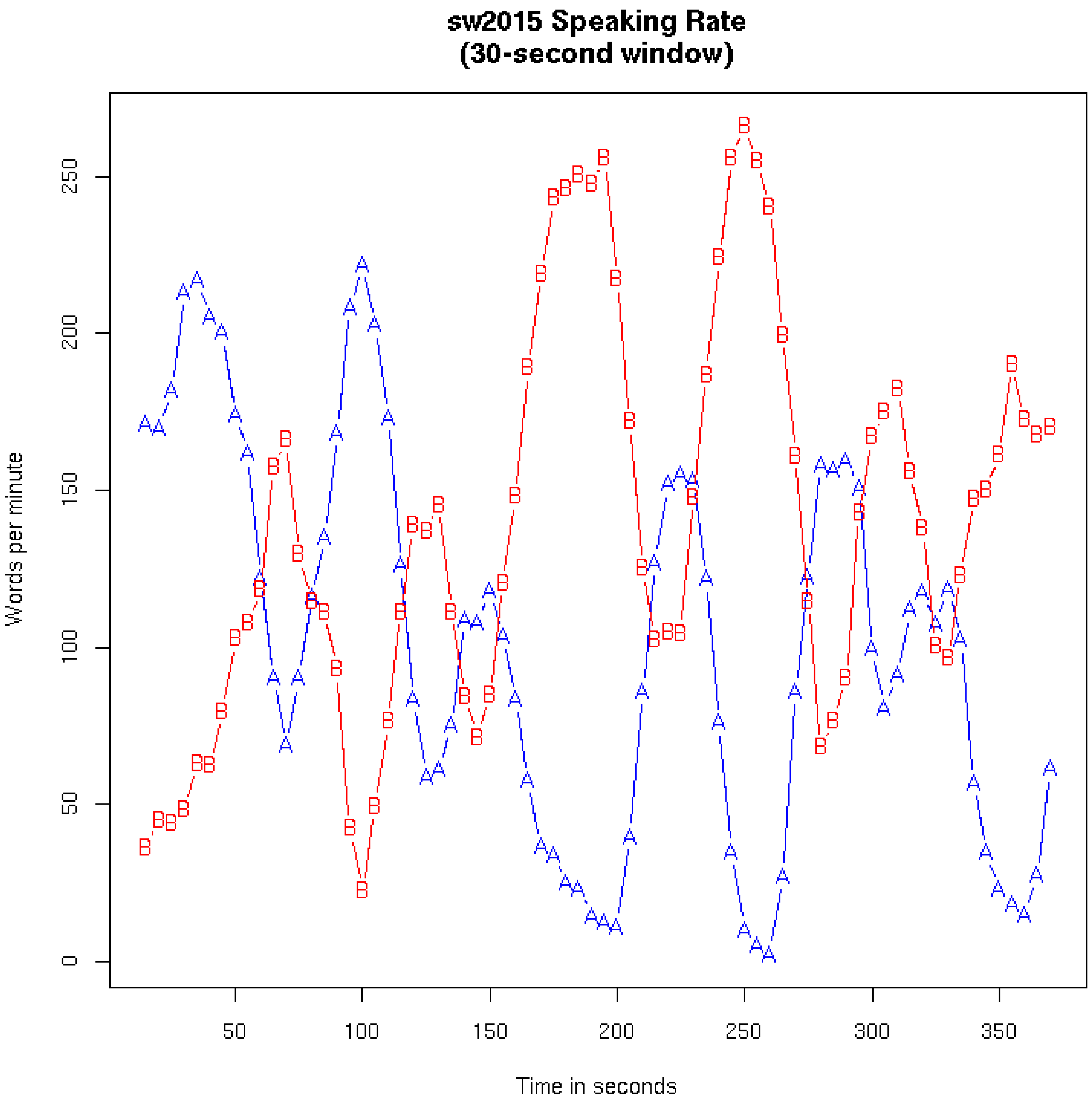
4. Does declination slope vary with phrase length?


eng_Bottom
man_Bottom
spa_Bottom
eng_Top
man_Top
spa_Top
Semitones
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
.0
.2
.4
.6
.8
1.0
Relative position in utterance


Slope (Semitones/second)
-2.0
-4.0
-6.0
-8.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Duration (Seconds)
Eng
Man
Spa

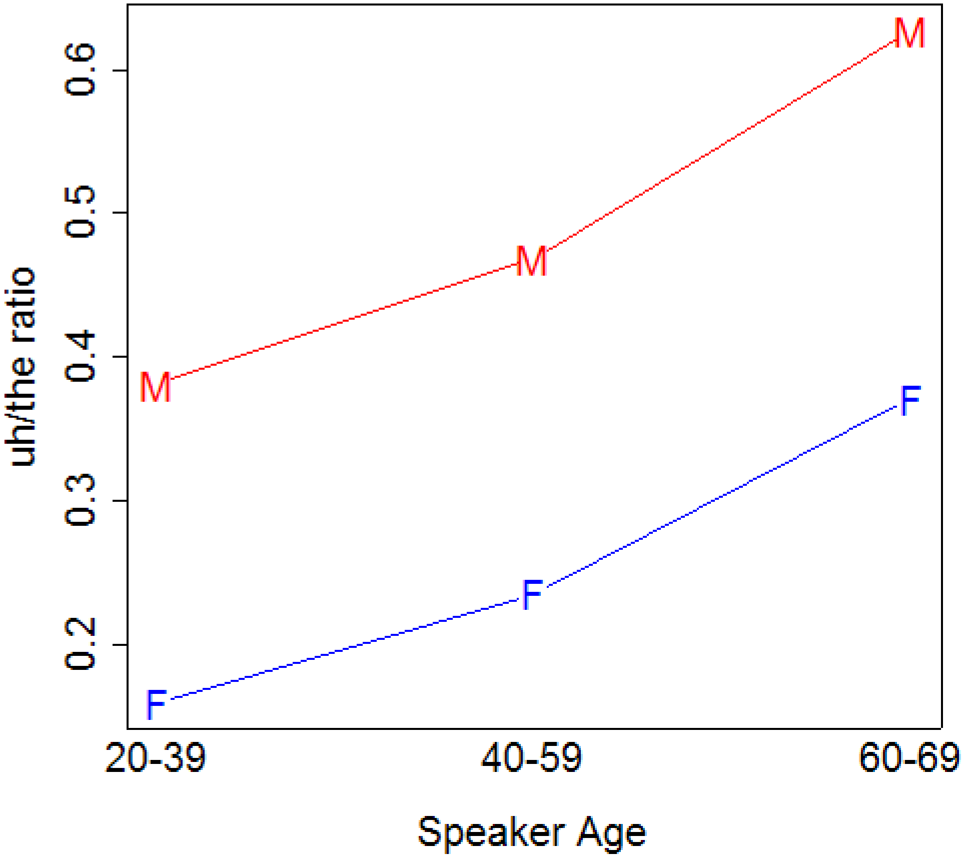

Mandarin Chinese
full
middle
Slope (Semitones/second)
-1.00
-1.50
-2.00
-2.50
-3.00
-3.50
-4.00
2.0
2.5
3.0
3.5
4.0
Duration (Seconds)

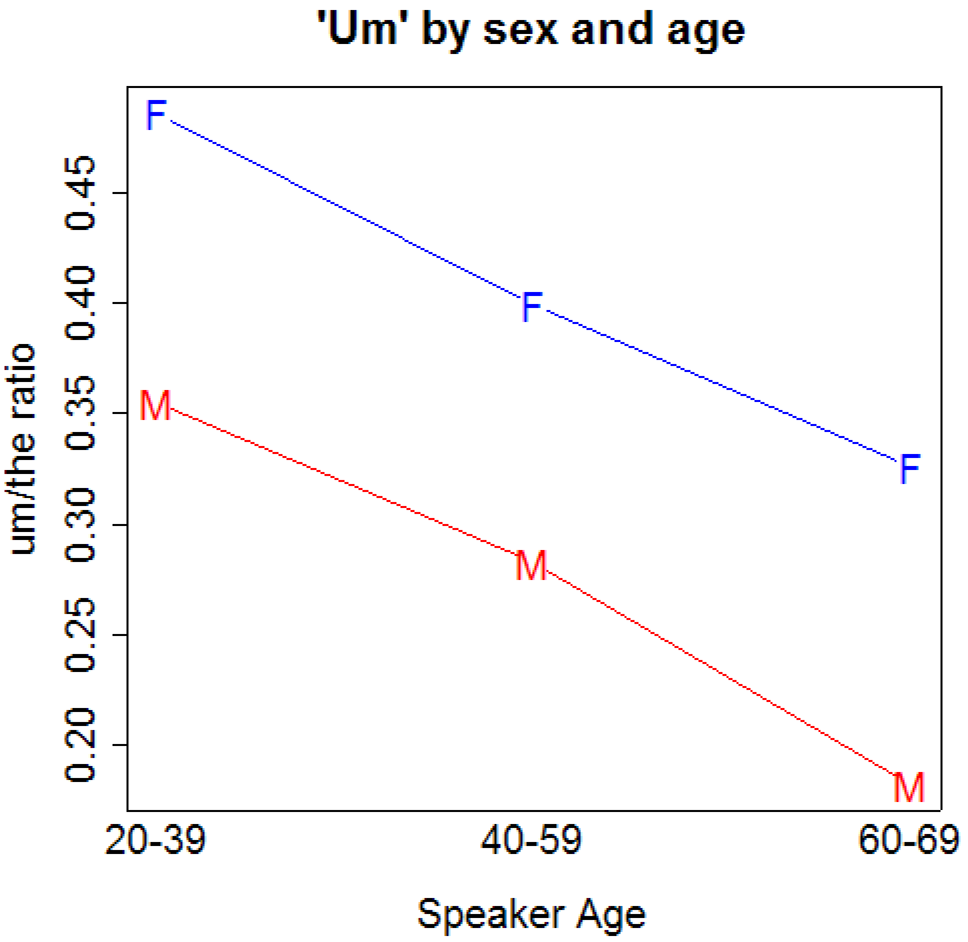
5. The shape of speech rate in a conversational interaction


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

6. Use of filled pauses by sex and age

'Uh' by sex and age





'Um' by sex and age

Enriching education

- Many eScience questions
  about speech and language
  are easy for students in high school
  and even in elementary school
  to understand and to investigate.
- Perhaps this will be the basis
  for reversing the trend
  to abandon linguistic analysis
  in primary and secondary education.
- While also teaching statistics
  and scientific reasoning!

Serious methodological issues

- For example, in phonetics
  - Orthographically-transcribed natural speech is available in very large quantities
  - By applying
    - Forced alignment,
    - pronunciation modeling,
    - automated measurements,

    we get a new world of phonetic data, in almost unlimited quantities
  - But natural data is very non-orthogonal and automated measurements may be problematic.
- Similar problems arise in analysis of other modalities.

Solutions are out there

- For example, hierarchical regression rather than analysis of variance….
- But the eSciences of Speech and Language pose somewhat different problems than Language Engineering does.
- We need a broadly-based community effort to define and address the issues.

Applications to other disciplines

- The basic eScience of Speech and Language is central.
- But similar techniques apply everywhere that speech and language are involved as objects of study or as data sources:

Psychology, neurology, anthropology, sociology, law, medicine, etc.

Back to Antonio’s roots?

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

What would Antonio do?

- Be enthusiastic about the opportunities
- Bring researchers together
- Persuade funders to invest

Thank you!