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?

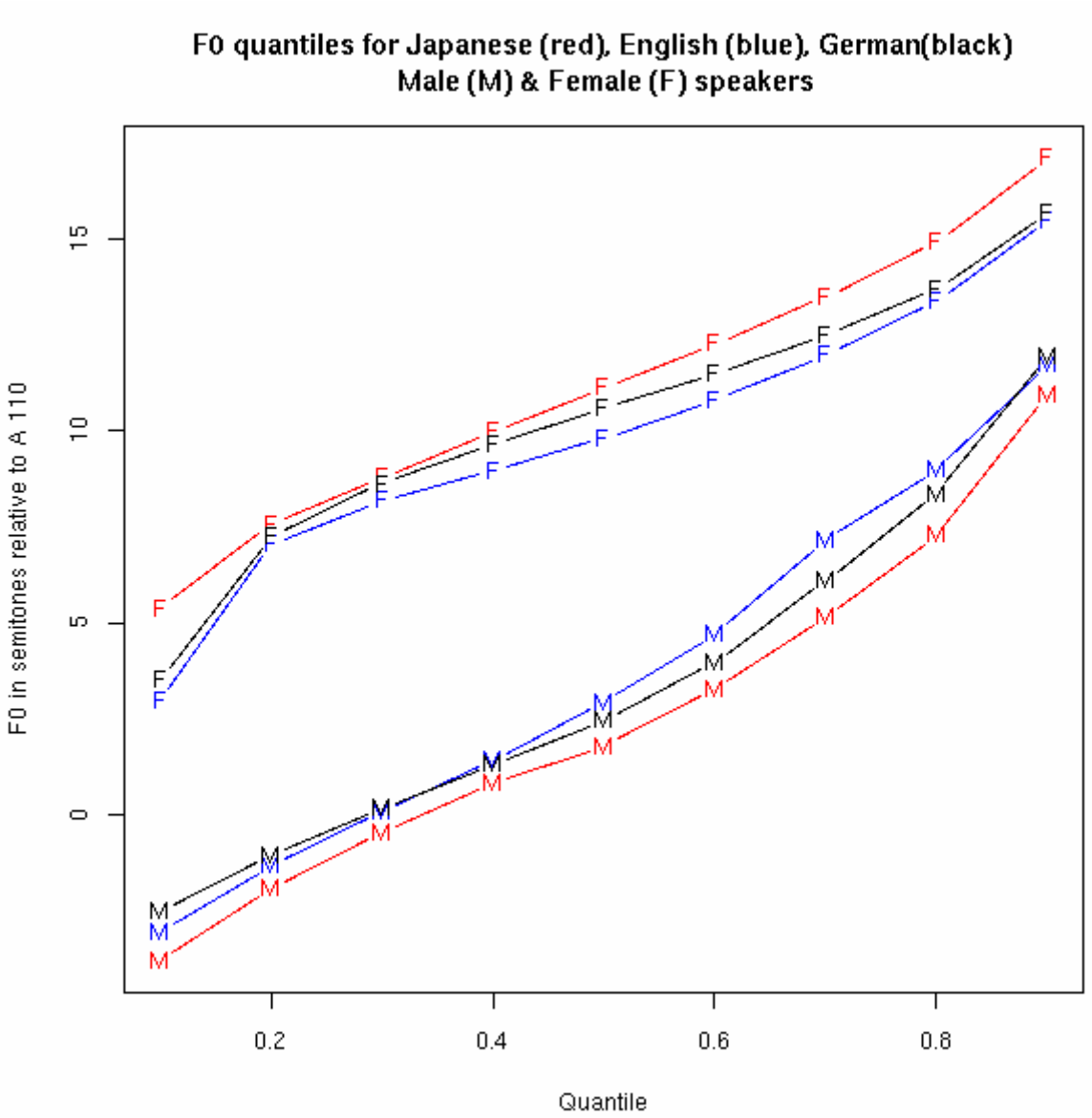
These are illustrative examples
 of questions that can be asked and answered in a few minutes
 with modern techniques and resources.

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.

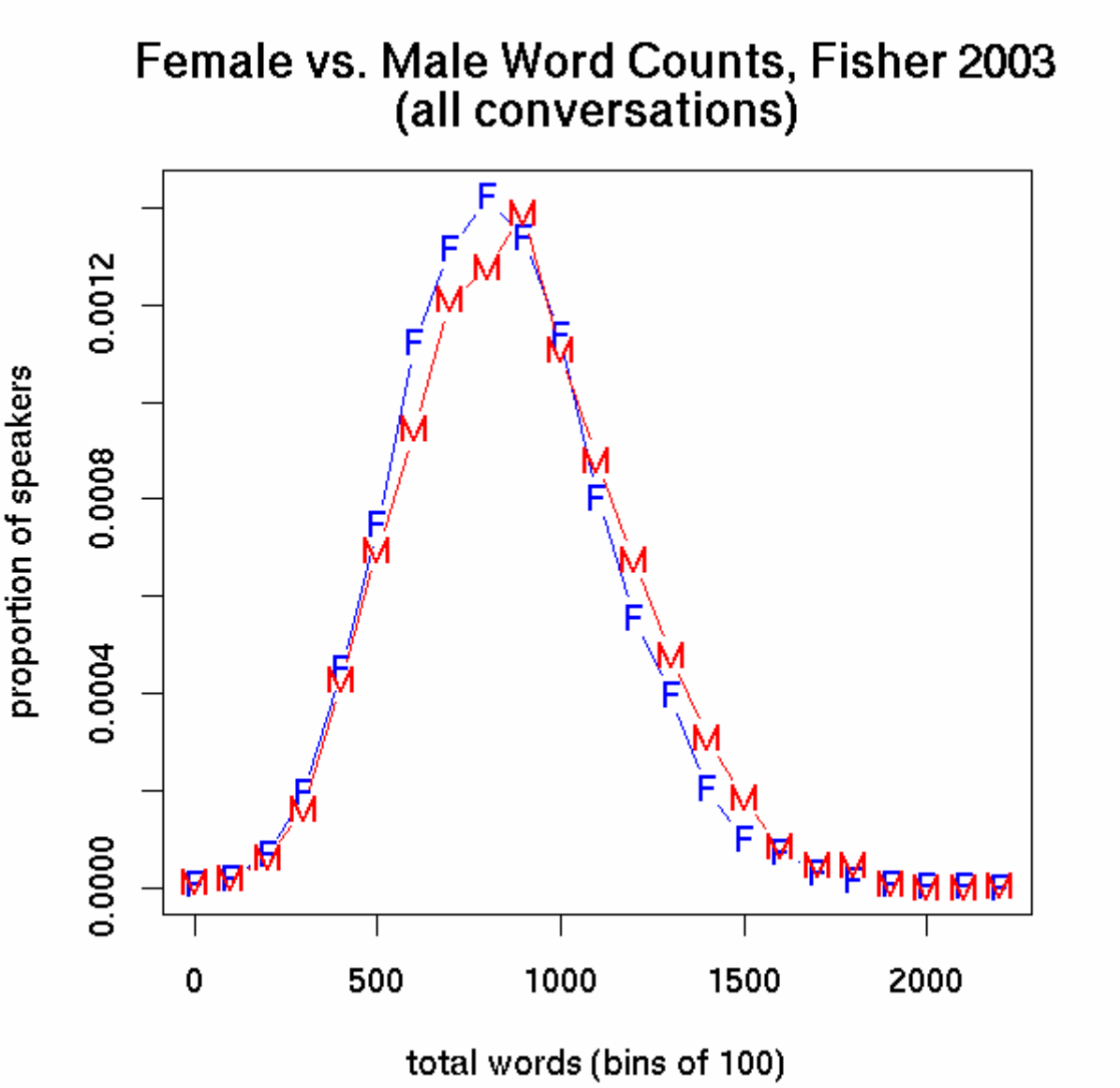
The point here
 is simply that interesting patterns are everywhere you look,
  and that large-scale looking
  has is now becoming increasingly easy.

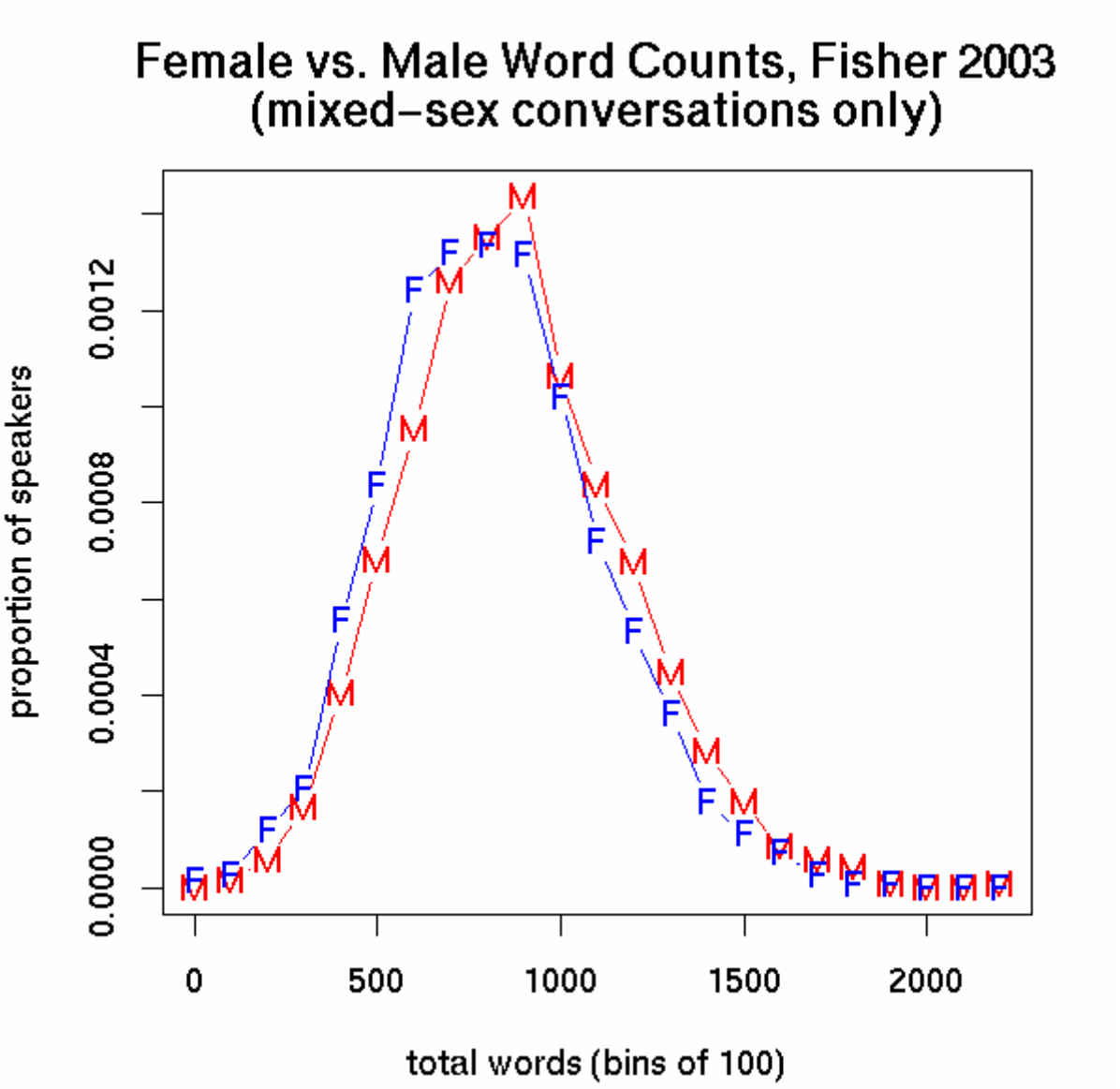
1. Cultural differences
   in gender polarization of pitch range



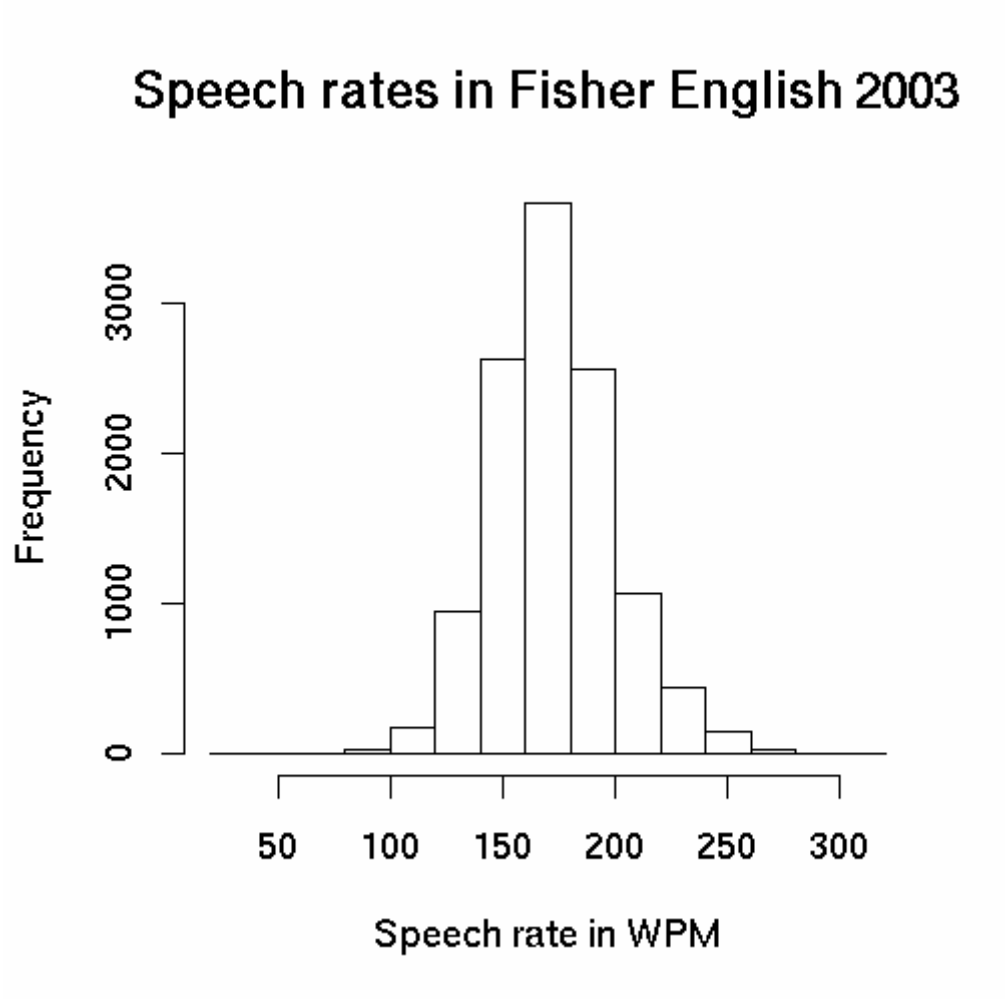
Data from CallHome M/F conversations; about 1M F0 values per category.

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


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

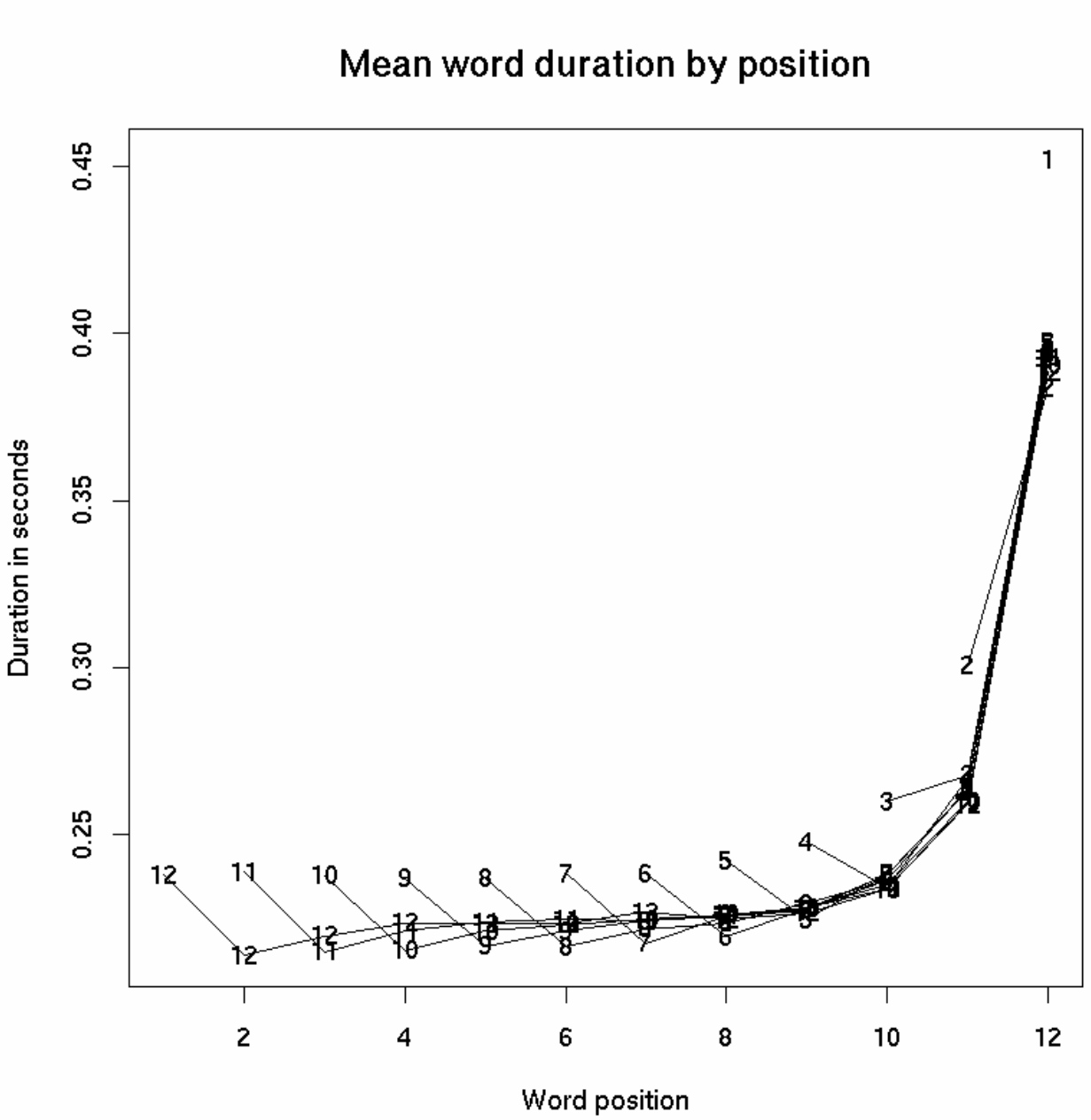

Female vs. Male Word Counts, Fisher 2003
(mixed−sex conversations only)
proportion of speakers
0.0000
0.0004
0.0008
0.0012
0
500
1000
1500
2000
total words (bins of 100)

2(b). Sex differences in speaking rate?

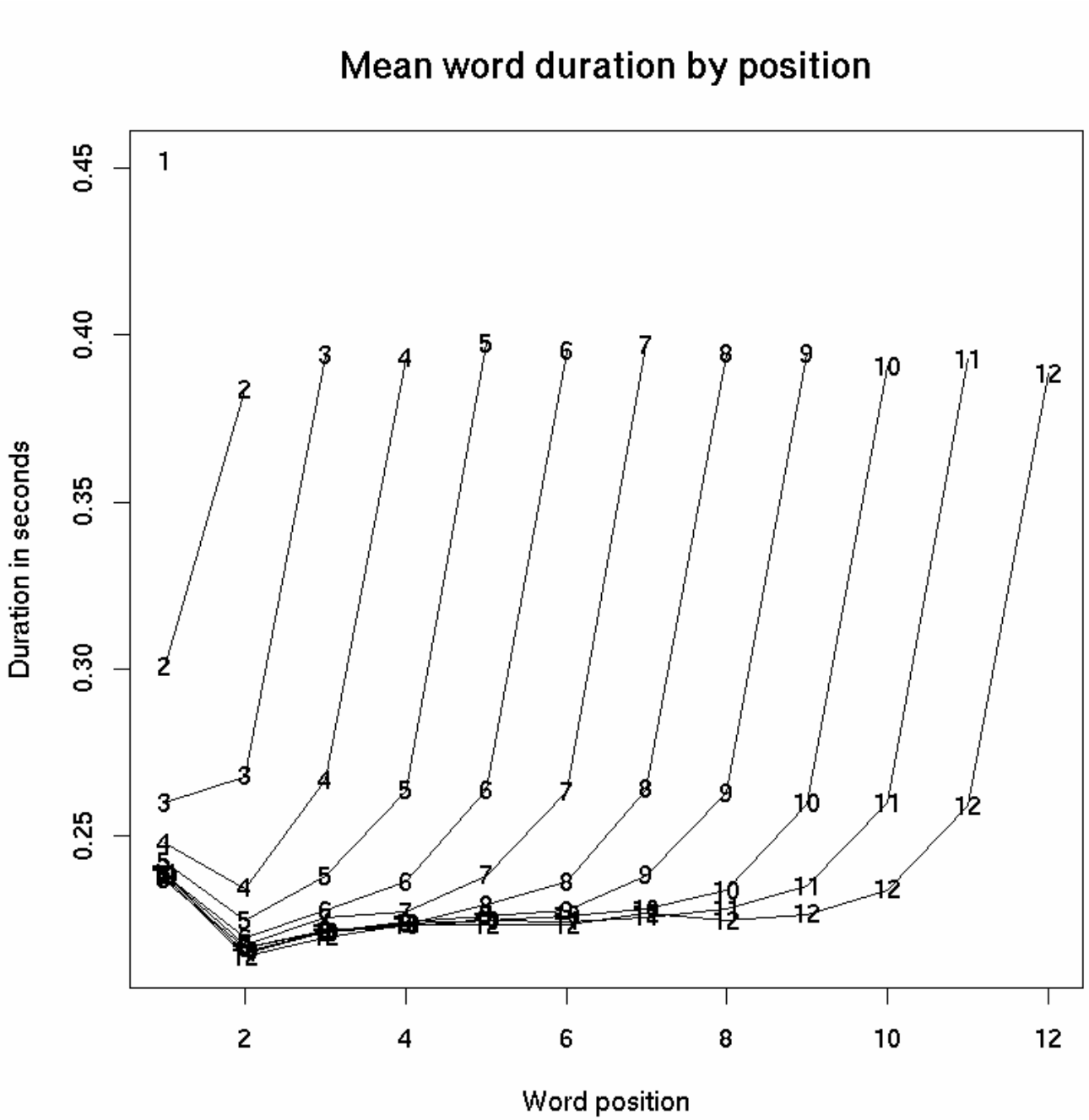


(11,700 conversational sides; mean=173, sd=27)
(Male mean 174.3, female 172.6: difference 1.7, effect size d=0.06)

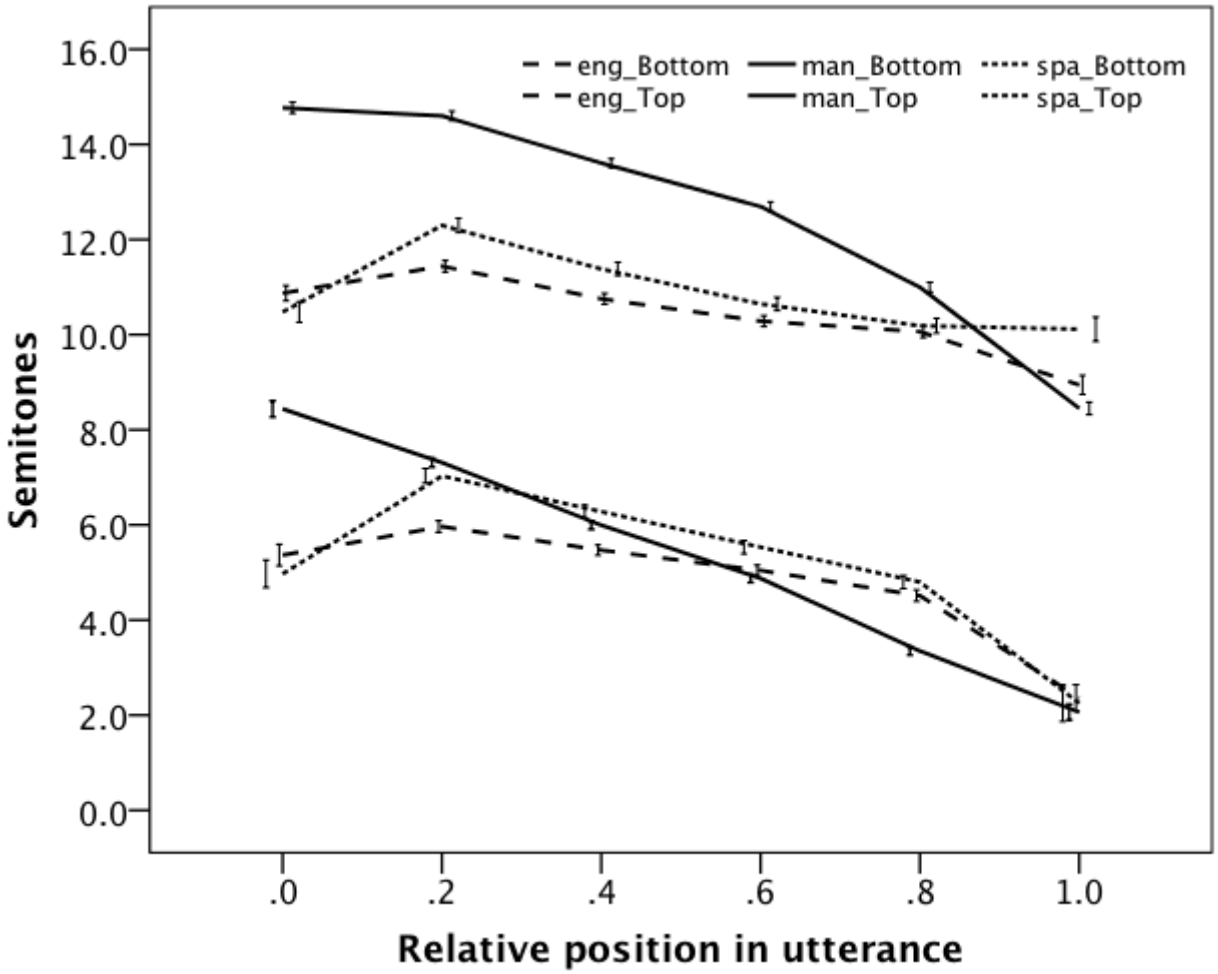
3. The shape of speech rate in a spoken phrase

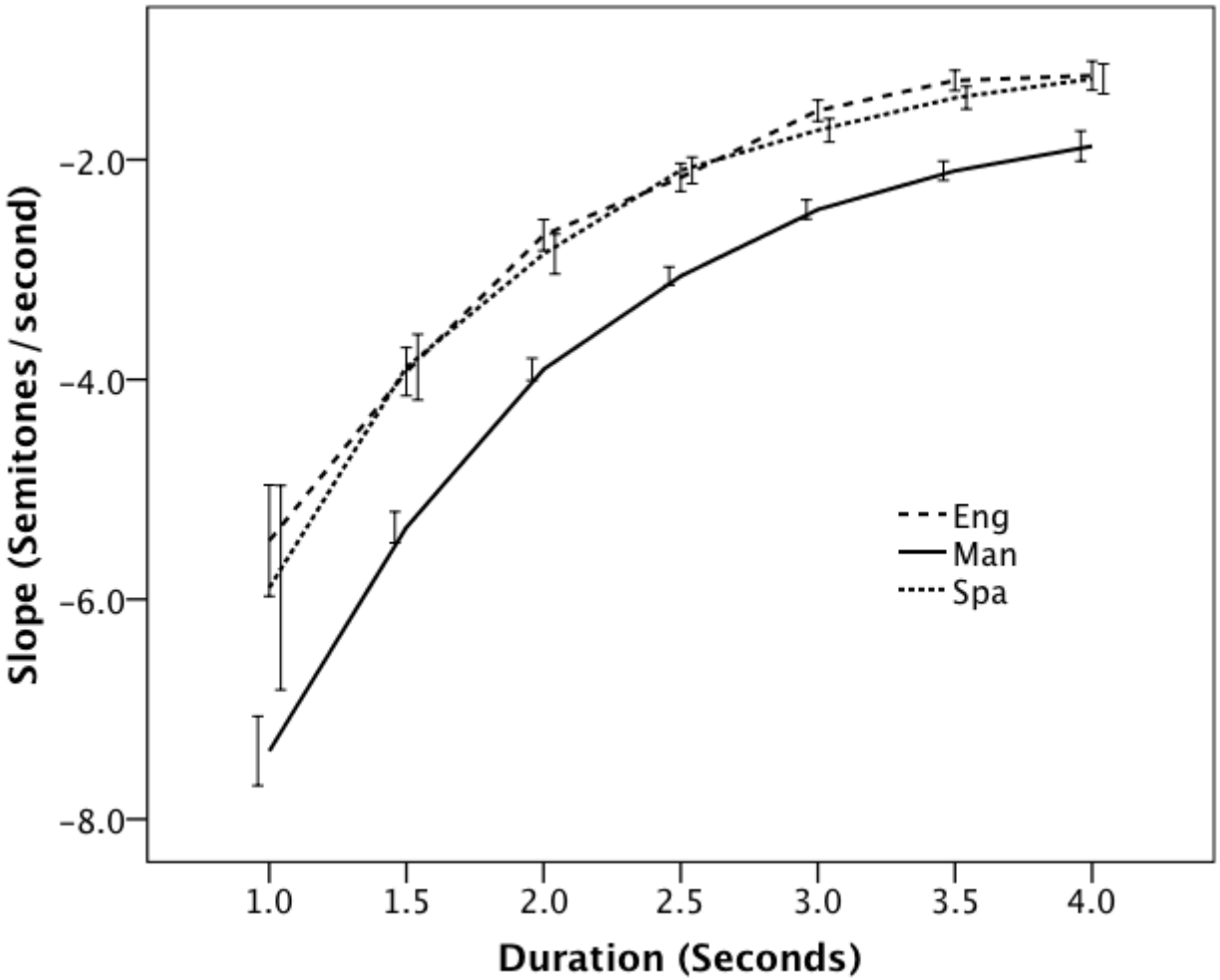


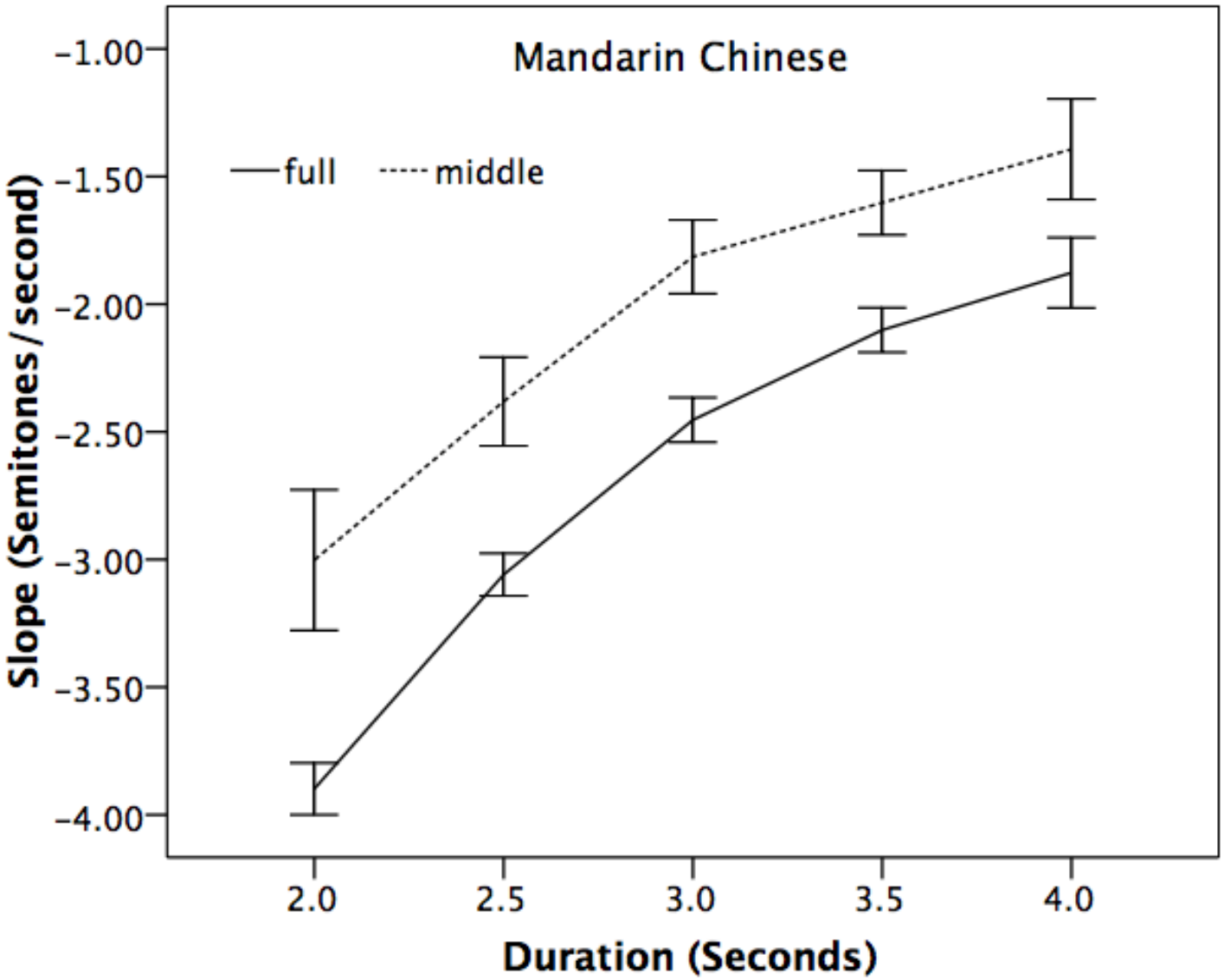
Data from Switchboard; phrases defined by silent pauses
(Yuan, Liberman & Cieri, ICSLP 2006)


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

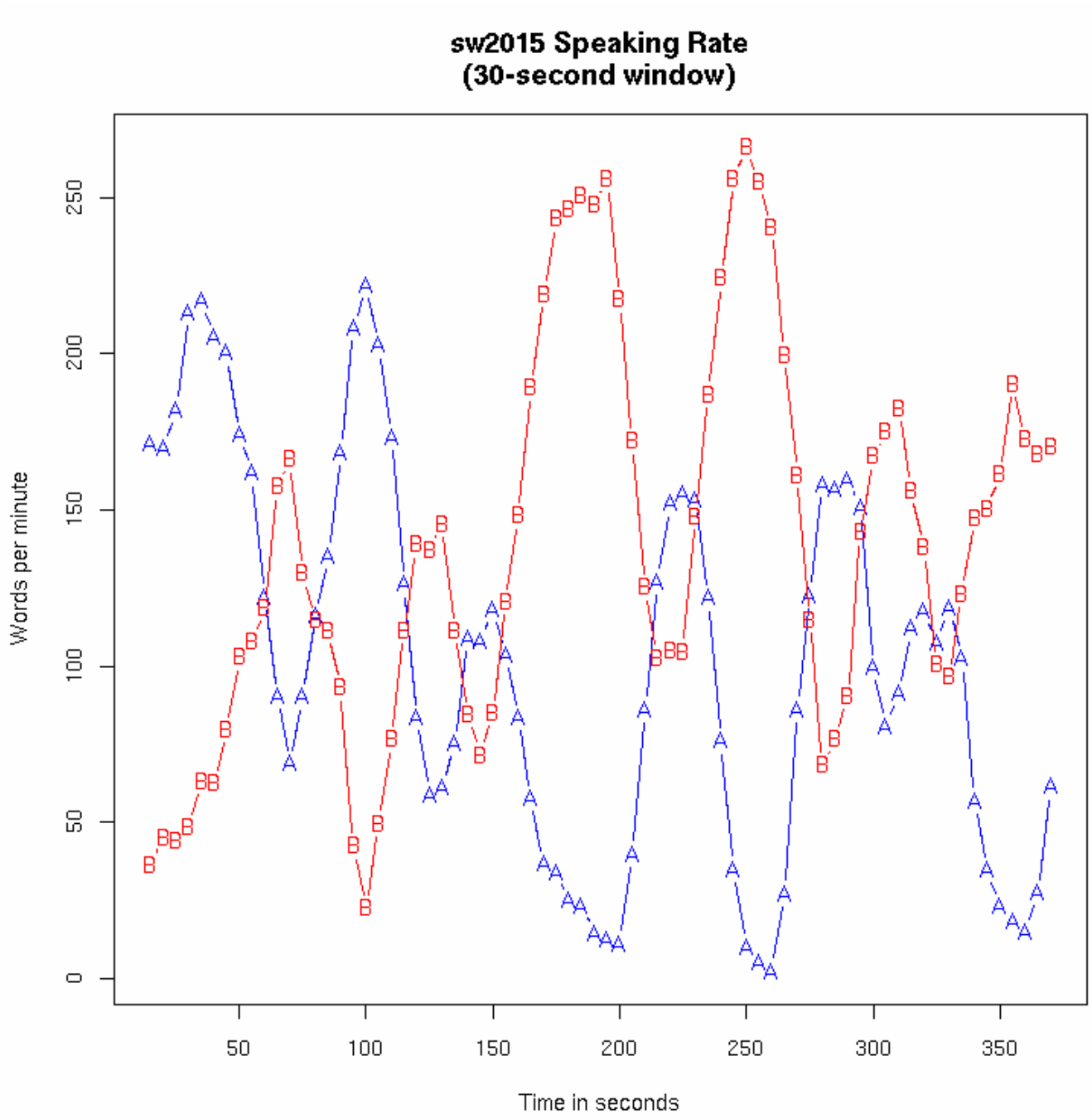
4. Does declination slope
vary with phrase length?


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

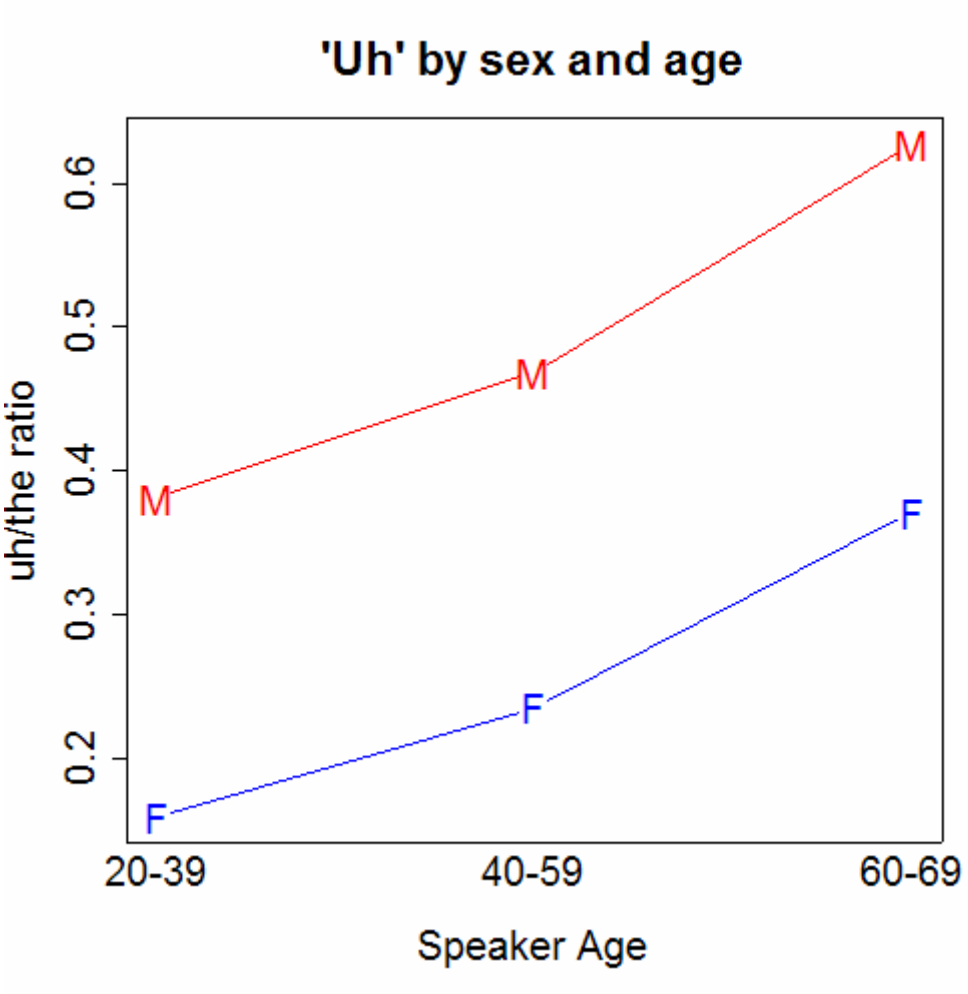

Slope (Semitones/second)
-2.0
-4.0
-6.0
-8.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Duration (Seconds)
Eng
Man
Spa

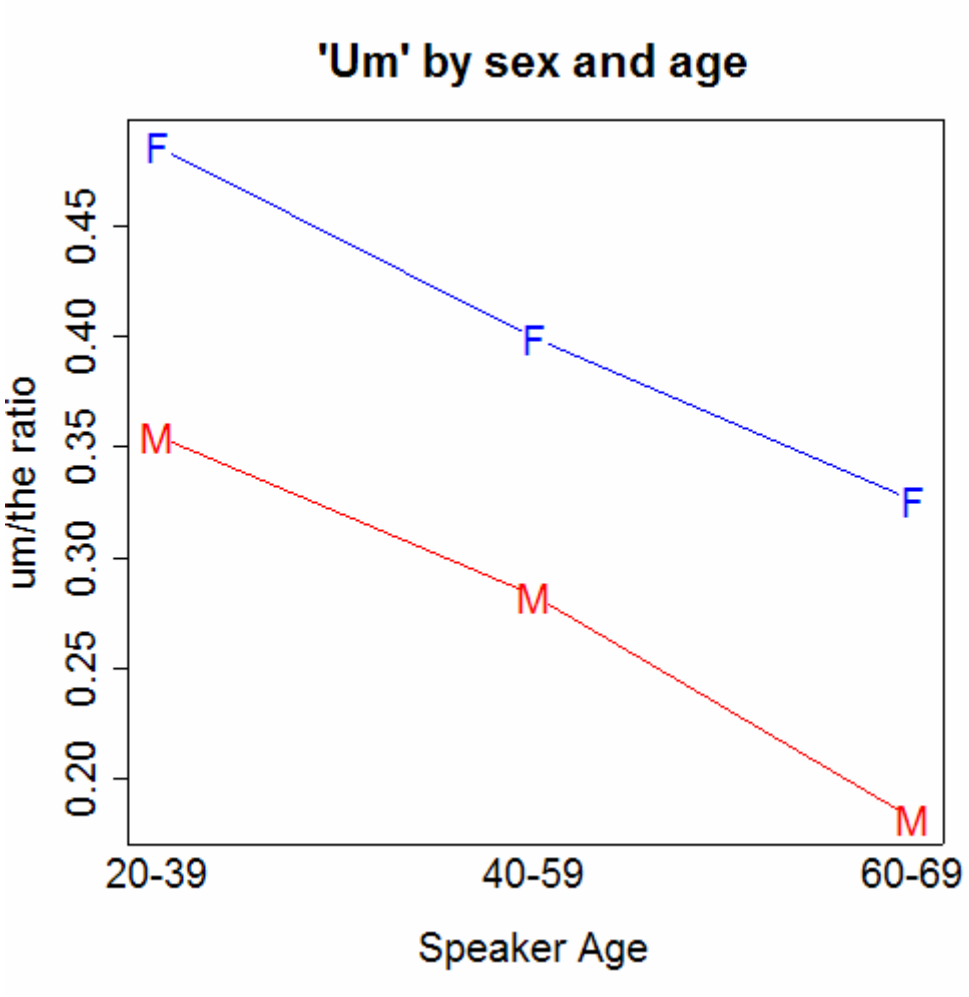

Mandarin Chinese
full
middle
-1.00
-1.50
-2.00
-2.50
-3.00
-3.50
-4.00
Slope (Semitones/second)
2.0
2.5
3.0
3.5
4.0
Duration (Seconds)

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

6. Use of filled pauses by sex and age


'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

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!

Another example

Noah Constant, Christopher Davis, Christopher Potts, and Florian Schwarz, “The pragmatics of expressive content: Evidence from large corpora” *Sprache und Datenverarbeitung*, 2009.

We use large collections of online product reviews, in Chinese, English, German, and Japanese, to study the use conditions of expressives (swears, antihonorifics, intensives). The distributional evidence provides quantitative support for a pragmatic theory of these items that is based in speaker and hearer expectations.

… and more …

Christopher Potts and Florian Schwarz, “Affective 'this‘”, *Linguistic Issues in Language Technology*, 2010.

Lakoff (1974) argues that affective demonstratives in English are markers of solidarity, with exclamative overtones deriving from their close association with evaluative predication. Focusing on this, we seek to inform these claims using quantitative corpus evidence. Our experiments suggest that affectivity is not limited to specific uses of *this*, but rather that it arises in a wide range of linguistic and discourse contexts. We also briefly extend our methodology to demonstrative *that* and to German *diese-* (‘this’).

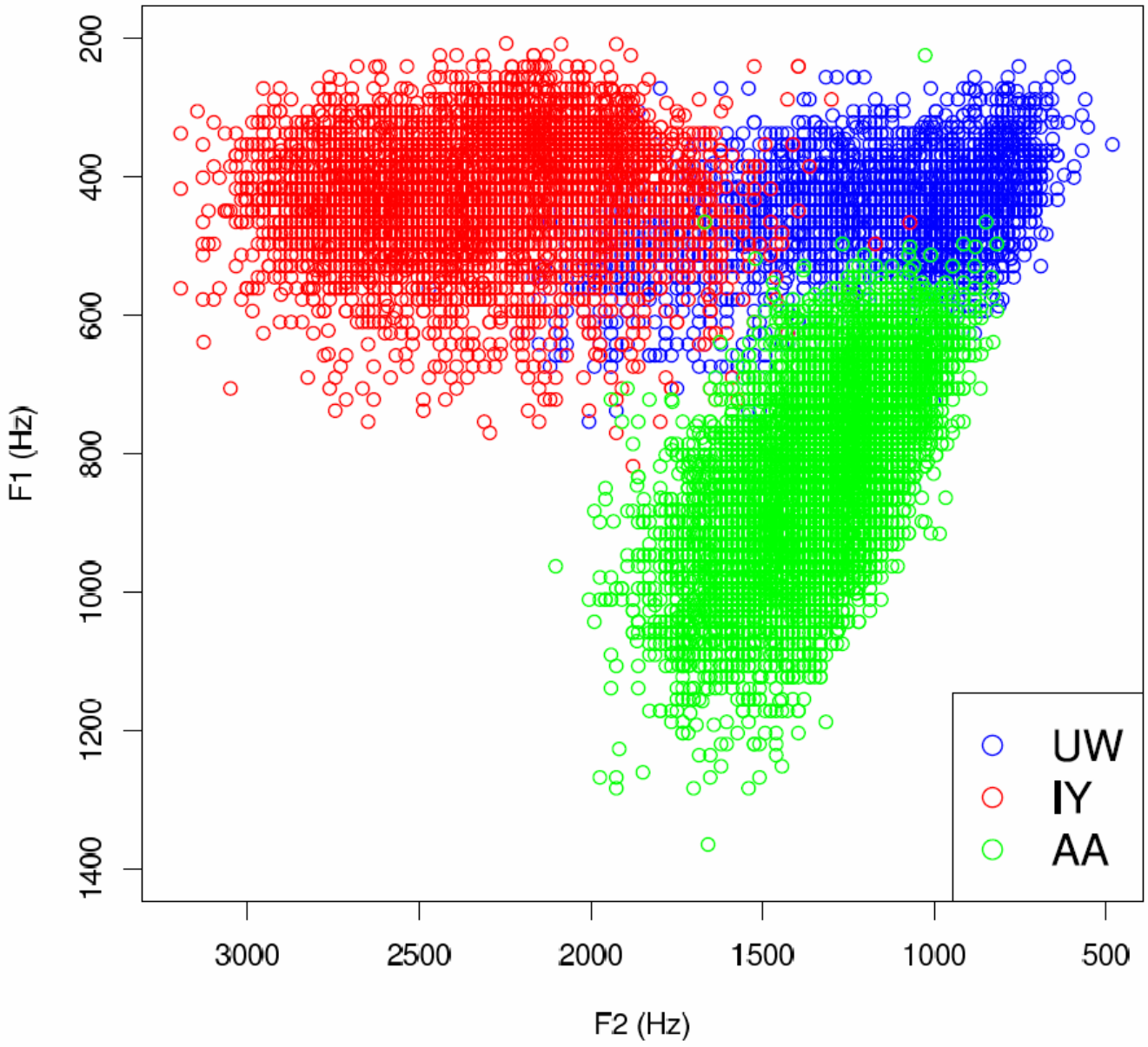
Automating sociolinguistic measurements

- W. Labov, S. Ash, and C. Boberg, *Atlas of North American English: Phonetics, Phonology and Sound Change*, Mouton 2005
- K. Evanini, S. Isard, and M. Liberman, “Automatic formant extraction for sociolinguistic analysis of large corpora”, *Interspeech* 2009
- K. Evanini, *The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania*, PhD diss. 2009.

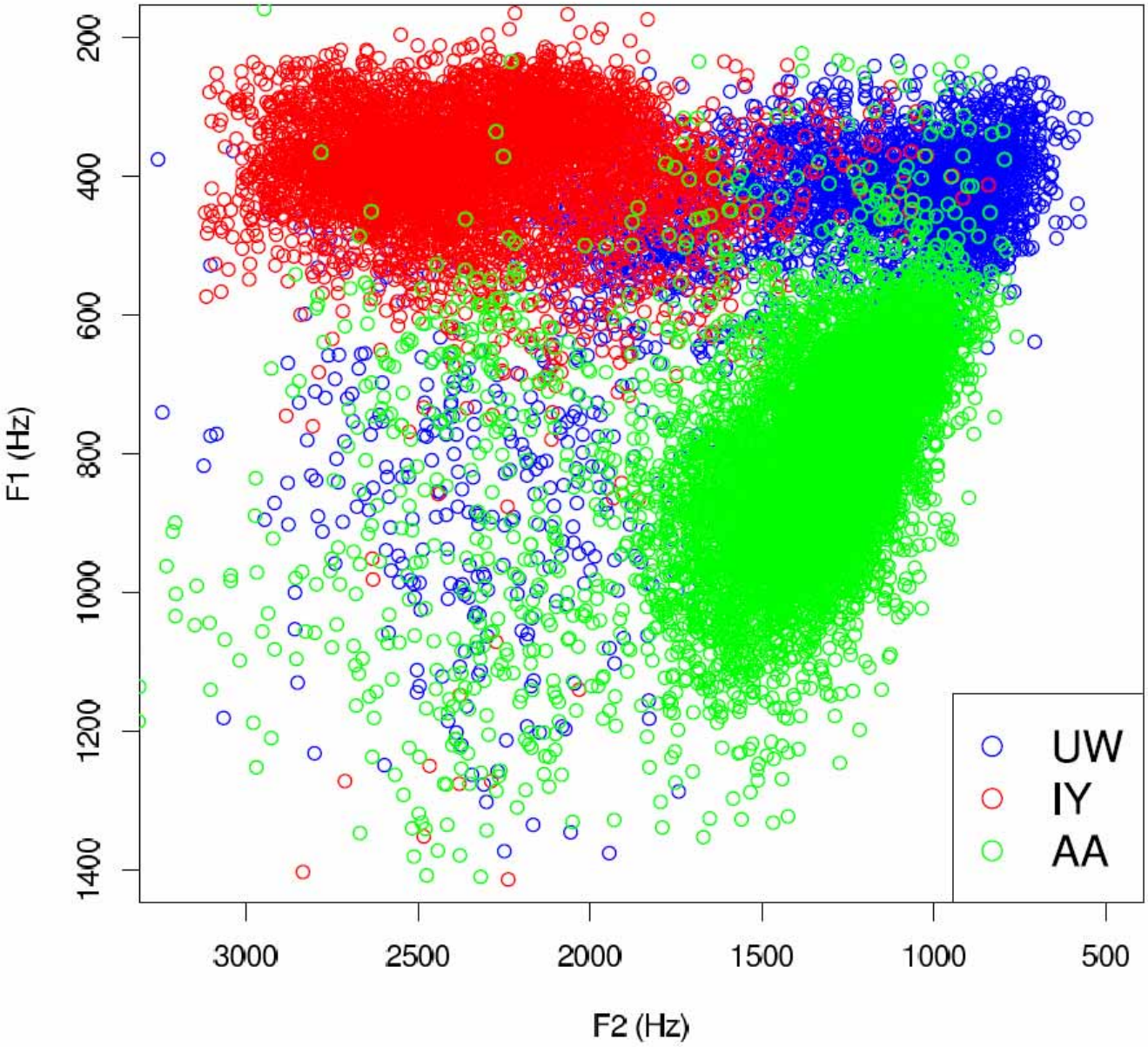
The ANAE corpus consists of ca. 30-minute long dialectological interviews conducted over the telephone with speakers from across the United States and Canada … at least two speakers were selected randomly from every city in North America with more than 50,000 inhabitants, and only speakers who had lived their entire lives in that city were chosen. […] A total of 439 speakers were selected for detailed acoustic
analysis by the ANAE authors. For these speakers, annotators examined all tokens with primary stress, and provided hand measurements for the first two formants at a single point in time.

For the purposes of comparing automatic formant prediction methods with the F1 and F2 values provided by the human ANAE annotators, it is necessary to determine the point in time at which the manual F1 and F2 measurements are taken. This information is not contained in the log files that were included in the published version of ANAE, but is available in earlier versions of the log files obtained from the ANAE authors. These two sources of information were merged to produce a database of F1 and F2 measurements with time stamps for a total of 111,810 tokens from 384 speakers (formant data from [the remaining] speakers had to be excluded because the original log files with time stamps were not available).

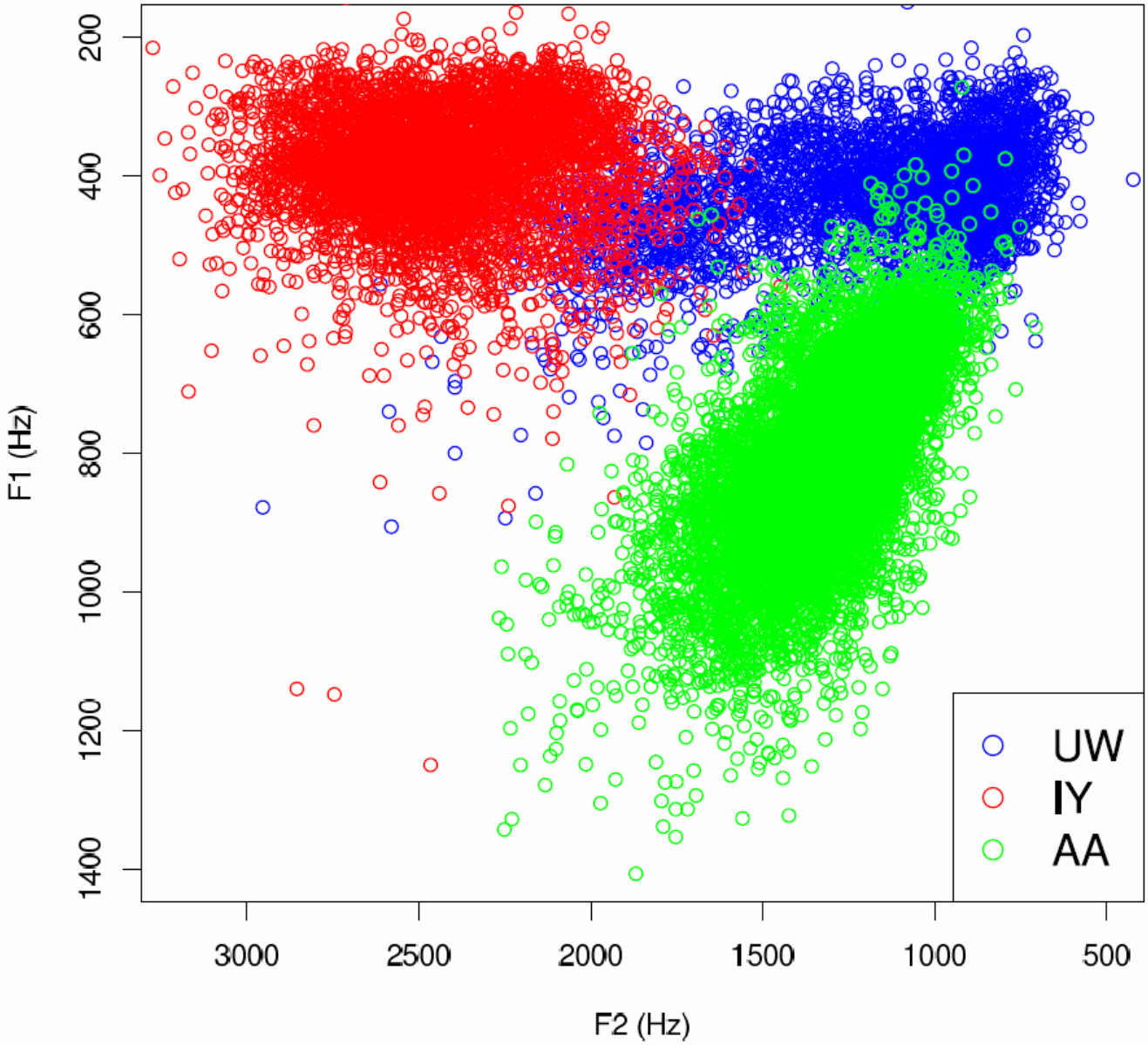
Manual Measurements



ESPS Formants



Proposed Method



Formant values	Accuracy
Manual	97.9%
Predicted (proposed method)	97.6%
ESPS default	90.2%

Table 2: Overall accuracy for classifying IY, UW, and AA using three different sets of F1 and F2 values (N = 17,954)

Evanini, Isard & Liberman, “Automatic formant extraction for sociolinguistic analysis of large corpora”, *Interspeech 2009*

Example 2:
Phonetics with real-world data

- Sproat & Fujimura 1993, Huffman 1997 showed that the allophonic difference between “clear” and “dark” [l] (in English) is a gradient one, depending on syllabic structure and speech rate.
- Yuan & Liberman 2009 replicated and extended these studies –

  using U.S. Supreme Court Justices as subjects.

Our Data

- The SCOTUS corpus includes more than 50 years of oral arguments from the Supreme Court of the United States – nearly 9,000 hours in total. For this study, we used only the Justices' speech (25.5 hours) from the 2001-term arguments, along with the orthographic transcripts.

- The phone boundaries were automatically aligned using the PPL forced aligner trained on the same data, with the HTK toolkit and the CMU pronouncing dictionary.

- This dataset contains 21,706 tokens of /l/, including

  3,410 word-initial [l]s,

  7,565 word-final [l]s, and

  10,731 word-medial [l]s.

Comparison

- Sproat & Fujimura 1993:
  ~200 tokens of /l/ in laboratory speech,
  hand measurement of formants
- Yuan & Liberman 2009:
  21,706 tokens of /l/
  from 2001 SCOTUS oral arguments
  automated clear/dark index
  based on fit to word-initial vs. –final models

Introduction

- Clear /l/ has a relatively high F_2 and a low F_1;
  Dark /l/ has a lower F_2 and a higher F_1;
  Intervocalic /l/s are intermediate between the clear and dark variants (Lehiste 1964).

- An important piece of evidence for the "gestural affinity" proposal: Sproat and Fujimura (1993) found that the backness of pre-boundary intervocalic /l/ (in /i - ɪ/) is correlated with the duration of the pre-boundary rime. The /l/ in longer rimes is darker.

- S&F (1993) devised a set of boundaries with a variety of strengths, to 'elicit' different rime durations in laboratory speech:

  Major intonation boundary: *Beel, equate the actors.* "|"

  VP phrase boundary: *Beel equates the actors.* "V"

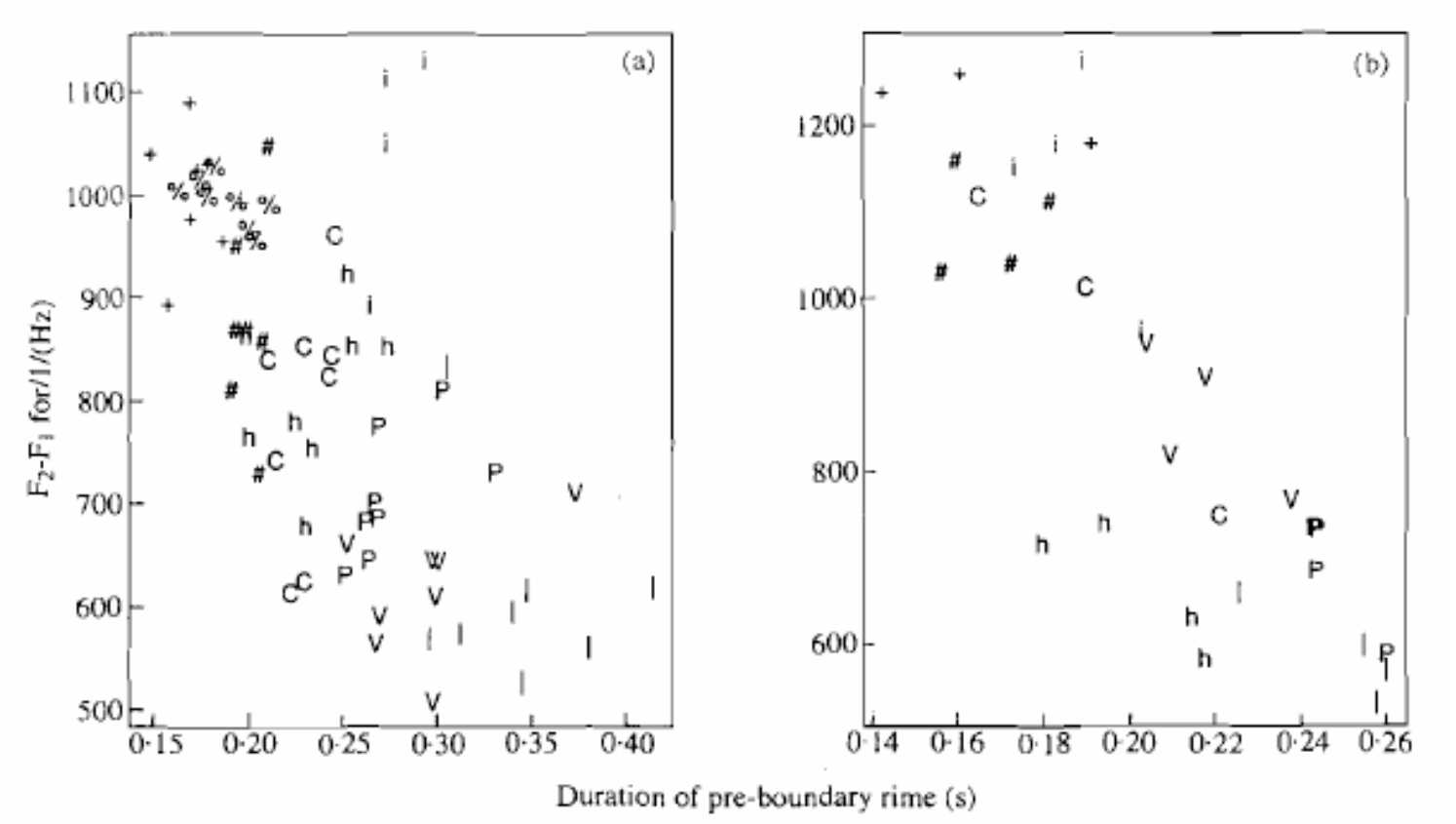
  Compound-internal boundary: *The beel-equator's amazing.* "C"

  '#' boundary: *The beel-ing men are actors.* "#"

  No boundary: *Mr Beelik wants actors.* "%"

Introduction

- Figure 1 in Sproat and Fujimura (1993):
  Relation between F_2-F_1 (in Hz) and **pre-boundary rime duration** (in s)
  for (a) speaker CS and (b) speaker RS.



"Forced Alignment"

- The aligner's acoustic models are GMM-based monophone HMMs on 39 PLP coefficients. The monophones include:

  speech segments: /t/, /l/, /aa1/, /ih0/, … (ARPAbet)

  non-speech segments:

  {sil} silence; *{LG}* laugh; *{NS}* noise; *{BR}* breath; *{CG}* cough; *{LS}* lip smack

  {sp } is a "tee" model with a direct transition from the entry to the exit node in the HMM (so "sp" can have 0 length)

  .... used for handling possible inter-word silence.

- The mean absolute difference between manual and automatically-aligned phone boundaries in TIMIT is about 12 milliseconds.

- http://www.ling.upenn.edu/phonetics/p2fa/

Method

- To measure the “darkness” of /l/ through forced alignment,
  we first split /l/ into two phones,
  L1 for the clear /l/ and L2 for the dark /l/,
  and retrained the acoustic models for the new phone set.

- In training, word-initial [l]’s (e.g., *like, please*) were categorized as L1 (clear); the word-final [l]s (e.g., *full, felt*) were L2 (dark).
  All other [l]’s were ambiguous, which could be either L1 or L2.

- During each iteration of training, the ‘real’ pronunciations of the ambiguous [l]’s were automatically determined,
  and then the acoustic models of L1 and L2 were updated.

- The new acoustic models were tested on both the training data and on a data subset that had been set aside for testing.
  During the tests, all [l]’s were treated as ambiguous –
  the aligner determined whether a given [l] was L1 or L2.

Method

- If we use word-initial vs. word-final as the gold standard, the accuracy of /l/ classification by forced alignment is 93.8% on the training data and 92.8% on the test data.

	L1	L2	
L1	2987	235	(training data)
L2	414	6757	
			⇐ gold-standard by word position
L1	169	19	
L2	23	371	(test data)

⇑

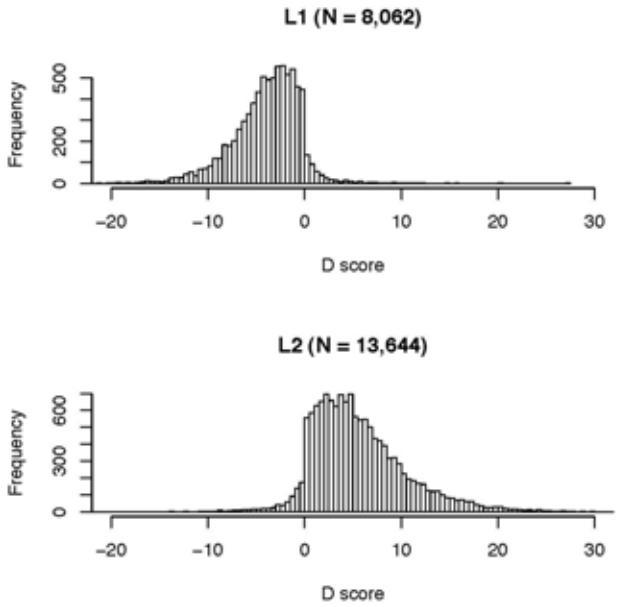
classified by the aligner

- *These results suggest that acoustic fit to clear/dark allophones in forced alignment is a plausible way to estimate the darkness of /l/.*

Method

- To compute a metric to measure the degree of /l/-darkness, we therefore ran forced alignment twice.
  All [l]'s were first aligned with L1 model, and then with the L2 model.
- The difference in log likelihood scores between L2 and L1 alignments – the *D score* – measures the darkness of [l].
  The larger the *D* score, the darker the [l].

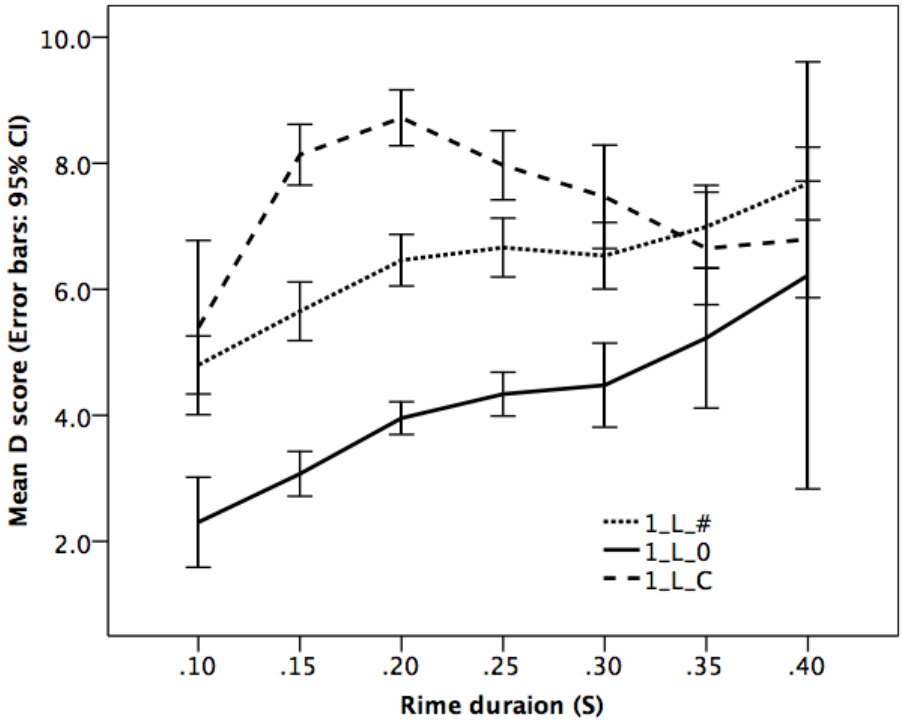
The histograms of the *D* scores:



Results

- To study the relation between rime duration and /l/-darkness, we use the [l]s that follow a primary-stress vowel (denoted as ‘1’).

- Such [l]s can precede a word boundary (‘#’), or a consonant (‘C’) or a non-stress vowel (‘0’) within the word.

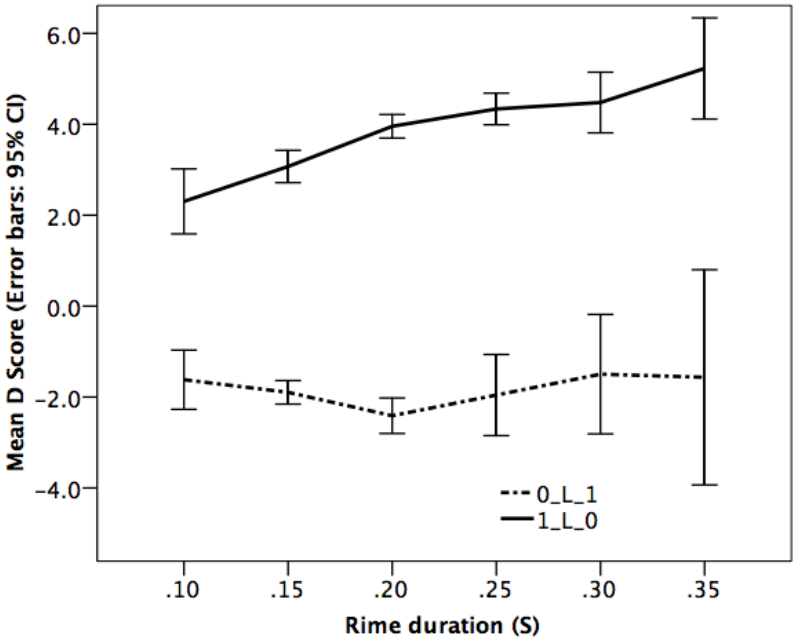


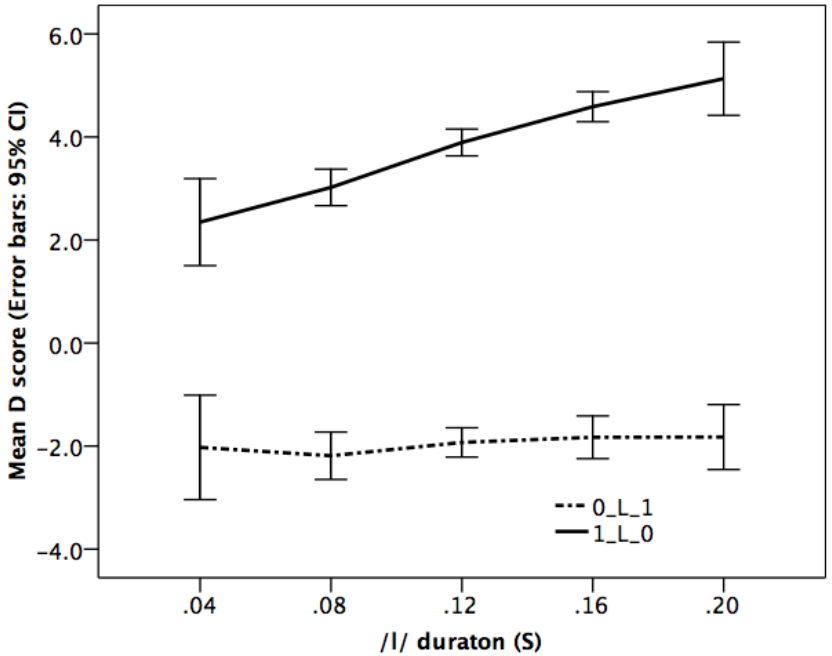
Yuan & Liberman: Interspeech 2009

Results

- To further examine the difference between clear and dark /l/,
  we compare the intervocalic (1_L_0) – syllable-final or "ambisyllabic"
  with the intervocalic (0_L_1) - syllable-initial

- The “rime” duration here means
  the duration of the previous vowel plus the duration of [l]
  regardless of putative syllabic affinity….





Yuan & Liberman: Interspeech 2009

Conclusions

?

Re-Inventing the Humanities

ePhilology: language documentation, language history, etc.

eRhetoric

ePoetics

eMusicology