The Golden Age of Speech and Language Science

Mark Liberman
University of Pennsylvania

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

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. Scientific use of these new instruments remains mainly potential, but the next decade is likely to be a new "golden age" of research. This talk will discuss some of the barriers to be overcome, present some successful examples, and speculate about future directions.
According to the 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.

*Language and Machines: Computers in Translation and Linguistics*

Report by the Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences
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.

Two wrinkles

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

... wait, what?

(2) And, the ALPAC report came out in 1966 (!) so 44 years later, where’s the QCD of 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
  2. engineering → new language science (???)

10/4/2010
Parenthesis: *What is linguistics?*

The term *linguistics* is ambiguous:

1. “rational inquiry into questions of speech and language”
2. “the institutions of academic linguistics”
3. “what people identified as linguists do”

In sense (1), most linguistics today is not done by linguists, but rather by computer scientists, electrical engineers, psychologists, neurologists, anthropologists, biologists, lexicographers, classicists, etc.

Still, . . .
Speech and language technology 1966-2010

• Despite the “funding winter” that followed ALPAC (and Pierce 1969), there has been an extraordinary flowering of technologies for dealing with digital speech and text.

• As more and more of the world’s intellectual and social life is mediated by digital networks, we can expect this to continue.
The science of speech 1966-2010

• Plenty of computer use
  – minicomputers in the 1960s
  – micro- and super-computers in the 1980s
  – ubiquitous laptops today
• Applications:
  – replaced tape splicing
  – replaced sound spectrograph
  – easier pitch tracking, formant tracking
  – more convenient statistical analysis
  – and so on
• BUT...
No phonetic quantum mechanics

• Great speech science by smart people
• But surprisingly little change
  – in style and scale of research 1966-2009
  – in scientific questions about speech
  – in the rate of progress compared to 1946-1966
    (the first golden age of phonetics)
    ...at least on the acoustic analysis side.
• Peterson & Barney 1951
  – data is still relevant
  – many contemporary publications
    are similar in style and scale
The science of language 1966-2010

• Despite ALPAC’s 1966 predictions, computers have not had a large impact on “the new linguistics” (in the sense of what “linguists” do).
• This has begun to change over the past decade, at least in some sub-disciplines.
• But for the most part, the effects are still in the future tense.
What went wrong?

• There are still many unmet challenges,
  – not enough new insights,
  – and the scientific potentialities of 1966 are still mostly potential.

• Is this just cultural conservatism?

• No. (1966-era) Computers were not enough: we also need
  – adequate accessible digital data
  – tools for large-scale automated analysis
  – applicable research paradigms

• Now: we have (at least) two out of three . . .
Why 2010 is like 1610

... at least a little ...

• Telescope: invented 1608
  Galileo 1609, Kepler 1611, Newton 1668
• Microscope: invented 1590
  Hooke 1665, Leeuwenhoek 1674

Instruments that opened new worlds to view
Why 2010 is like 1610

Science needs theory -- but

“Sometimes you can observe a lot just by watching”

-Yogi Berra
Breakfast experiments

• Our “telescope and microscope” are
  – Easily available collections of speech and text
  – Computer algorithms for
    • analyzing speech and text
    • aligning 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.
Five One-Hour Explorations

• 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?
• How does disfluency vary with sex and age?
One-hour exploration #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
    • about 100 conversations per language
  – Compare quantiles of pooled values
• Answer: yes, apparently so.
Data from CallHome M/F conversations; about 1M F0 values per category.

10/4/2010

CUHK: The Golden Age
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
One-hour exploration #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)

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)
One-hour 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.
Speech rates in Fisher English 2003

(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)
One-Hour 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

Word position

10/4/2010

CUHK: The Golden Age
One-hour 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)
One-hour experiment #5

• How does disfluency vary with sex and age?
• Method: count “filled pauses” in transcripts of U.S. English conversations by demographic categories of speakers
• Result: systematic but unexpected interaction
'Uh' by sex and age

uh/the ratio

Speaker Age

20-39 40-59 60-69

M M M

F F F

CUHK: The Golden Age
'Um' by sex and age

Speaker Age

um/the ratio

20-39  40-59  60-69

F

M

F

M
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.

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!
Thank you!
Serious example #1: Automating sociolinguistic measurements


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).
ESPS Formants

![Diagram showing ESPS Formants with F1 and F2 axes. The diagram includes clusters of points differentiated by color: UW (blue), IY (red), and AA (green).](image-url)
Proposed Method

![Chart showing data points with labels UW, IY, and AA.](chart.png)
Table 2: Overall accuracy for classifying IY, UW, and AA using three different sets of F1 and F2 values (N = 17,954)

<table>
<thead>
<tr>
<th>Formant values</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>97.9%</td>
</tr>
<tr>
<td>Predicted (proposed method)</td>
<td>97.6%</td>
</tr>
<tr>
<td>ESPS default</td>
<td>90.2%</td>
</tr>
</tbody>
</table>

Evanini, Isard & Liberman, “Automatic formant extraction for sociolinguistic analysis of large corpora”, *Interspeech 2009*
Serious example #2: Exploring patterns of pitch

“Declination” is the tendency for voice pitch to fall, on average, over the course of a phrase.

Studied quantitatively for more than 50 years.

What happens when we point our new instruments in this direction?

Jiahong Yuan & Mark Liberman, “F₀ Declination in English and Mandarin Broadcast News Speech”, InterSpeech 2010.
Introduction (1)

- Declination refers to the downward trend of $F_0$ over the course of an utterance (Cohen, Collier and 't Hart, 1982).

- Possible physiological causes for $F_0$ declination:
  - Downtrend of subglottal pressure (Lieberman, 1967)
  - Tracheal pull (Maeda, 1976)

- Debates about $F_0$ declination:
  - Is it (only) a by-product of the physics and physiology of talking? Or is it (also) a linguistically-controlled feature? (Ladd 1984, Strik and Boves 1995).
  - Is it (only) the result of local events and fixed-rate processes? Or does it require phrase-scale pre-planning?
Introduction (2)

Findings and theories about F₀ declination:

- The slope of the declination depends on the sentence type: it is steepest for terminal declaratives and least steep for interrogatives (Thorsen, 1980).
- The initial F₀ peak increased with sentence length (Cooper and Sorensen, 1981).
- Declination could be explained by (local) downstep plus (local) final lowering effects (Liberman and Pierrehumbert, 1984).
- The F₀ topline, i.e., the line connecting F₀ peaks, was steeper in short sentences than in long ones (Swerts et al., 1996).
- Final lowering is grammaticalized and independent of declination (Arvaniti, 2009).
Introduction (3)

• Besides declination, the $F_0$ contour of a sentence is affected by many linguistic and situational factors.

• Most previous studies of declination used controlled experiments to minimize the effects of the other factors.

• In this study, we investigate $F_0$ declination in large broadcast news speech corpora.

• We hypothesize that in large natural speech corpora, the other factors will balance and cancel each other, and the “true” declination effect will be revealed.

• In any case, this source of evidence allows us to examine the debates about declination from another angle.
Introduction (4)

- \( F_0 \) declination is also found in Mandarin Chinese (Xu, 1999):
  - Declination on high tone sequences (Shih 2003)
  - Linear baseline declination (Yuan 2004)

- We will compare the declination effect in English and in Mandarin Chinese.
Data

• Three broadcast news speech corpora were used:
  – 1997 English Broadcast News Speech (LDC98S71)
  – 1997 Mandarin Broadcast News Speech (LDC98S73)
  – 1997 Spanish Broadcast News Speech (LDC98S74)

• The “utterances” (between-pause units in the transcripts) were aligned with the transcripts using the PPL Forced Aligner:
  – Units containing a pause longer than 50 ms were excluded.
  – Units from unknown speakers were also excluded.

• Utterances 1-4 seconds long were selected for this study:
  – 5,652 in English; 8,383 in Mandarin; 3,904 in Spanish

(All data and scripts will be published in association with the paper)
Methods (1)

• $F_0$ contours were extracted using `esps/get_f0` with a 10 ms frame rate.

• The contours were linearly interpolated to be continuous over the unvoiced segments, and smoothed by passing them through a Butterworth low-pass filter with normalized cutoff frequency at 0.1.

• $F_0$ values were converted to semitones. The base frequency used for calculating semitones was speaker dependent, defined as the 5th percentile of all $F_0$ values for that speaker.

\[
\text{Semitone} = 12 \times \log_2\left(\frac{F_0}{F_0_{\text{base}}}\right)
\]
Method

- Two methods were applied to measure $F_0$ declination.
  - A linear **regression line** was fitted to each contour using the least-squares method.
  - The convex-hull algorithm (Mermelstein, 1975) was used to identify local $F_0$ valleys and peaks. The $F_0$ peaks were connected to form the **topline** and the $F_0$ valleys were connected to form the **baseline**.
Results: Length and Slope

- There is a strong correlation between utterance length and declination slope: The shorter the utterance, the steeper the slope of the regression line is.
And also in Spanish!

Our Spanish data were very close to English:
Results: Length and Slope, Excluding Edges

- The time-slope relationship still holds after excluding the initial and final 500 ms. That is, shorter utterances have steeper declination within the middle region, from which likely “edge effects” are excluded.
Results: Topline and Baseline

• Both the topline and baseline show declination, and the topline has final lowering in both languages.
• The baseline of Mandarin Chinese is close to a straight line.
• In English the topline and baseline are very similar: initial rising, middle declination, and final lowering.
And also in Spanish!

- Again, Spanish is very much like English (in this measure, on this data):
Results: Pitch Range, Number of Extrema

- In this sample, Mandarin Chinese has wider pitch range than English.
- Mandarin Chinese also has more $F_0$ peaks and valleys than English, i.e. more $F_0$ fluctuations
Conclusions and discussion (1)

- We investigated $F_0$ declination in large broadcast news speech corpora in English and Mandarin Chinese (… and Spanish).

- We applied two methods, linear regression and convex hull. Regression was used to measure overall declination slope, and the convex hull was used to extract $F_0$ peaks and valleys for representing the topline and baseline patterns.

- There was a strong relation between declination slope and utterance length: the shorter the utterance, the steeper the declination. This relationship still holds when the first and last 0.5 sec. are excluded.

- This result suggests that speakers may control declination slope (directly or indirectly) in a way that requires phrase-level planning.
Conclusions and discussion (2)

- Both the topline and baseline show declination, and the topline has final lowering -- in all three languages.
- In Mandarin Chinese, the baseline is close to a straight line, which is different from the topline.
- In English and Spanish, the baseline and topline are similar, both consisting of three parts: initial rising, middle declination, and final lowering.
- This cross-linguistic difference suggests that topline and baseline declinations are partly independent phenomena, and that they are not entirely automatic by-products of some physiological process, but are linguistically controlled at least in part.
- Finally, our results showed that Mandarin Chinese has wider pitch range and more $F_0$ fluctuations than English.

This is probably due to the effect of lexical tone, but may also represent a difference in broadcast styles.
More on Sources

- 1997 English Broadcast News Speech (*LDC98S71*):
  ABC World News Tonight, CNN Headline News, CNN Early Prime,
  CNN Prime News, CNN The World Today, PRI The World,
  CSPAN Public Policy, CSPAN Washington Journal

- 1997 Mandarin Broadcast News Speech (*LDC98S73*):
  CC-TV, KAZN-AM, VOA

- 1997 Spanish Broadcast News Speech (*LDC98S74*):
  ECO, Univision, VOA
Serious example #3

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 ...


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').