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.

-Diane Coyle, “Economics on the Verge of a Golden Age”,
*The Chronicle of Higher Education*, March 12, 2010
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 (???)
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 much of an 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

Yogi Berra:

“Sometimes you can observe a lot just by watching”
Examples: “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?
• How does disfluency vary with sex and age?
F0 quantiles for Japanese (red), English (blue), German (black) Male (M) & Female (F) speakers

Data from CallHome M/F conversations; about 1M F0 values per category.
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)
(11,700 conversational sides; mean=173, sd=27)
(Male mean 174.3, female 172.6: difference 1.7, effect size d=0.06)
Data from Switchboard; phrases defined by silent pauses
(Yuan, Liberman & Cieri, ICSLP 2006)
Mean word duration by position

Duration in seconds

Word position

2 4 6 8 10 12

0.25 0.30 0.35 0.40 0.45
'Uh' by sex and age

 uh/the ratio

 Speaker Age

 20-39 40-59 60-69

 M M F

 M F F

 F F F
Serious corpus 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
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).
Proposed Method

![Graph showing proposed method with labeled axes and scattered data points in different colors representing UW, IY, and AA categories.](image-url)
Evanini, Isard & Liberman, “Automatic formant extraction for sociolinguistic analysis of large corpora”, *Interspeech 2009*

<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>

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

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>2987</td>
<td>235</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>414</td>
<td>6757</td>
<td></td>
</tr>
</tbody>
</table>

(training data)

\[ \Leftarrow \text{gold-standard by word position} \]

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>169</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>23</td>
<td>371</td>
<td></td>
</tr>
</tbody>
</table>

(test data)

\[ \uparrow \text{classified by the aligner} \]

- 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.
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….
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.
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’).
Re-uniting (some of) Snow’s two cultures?

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.

Conclusions