

# The story begins in the 1960s...

... with two bad reviews by John Pierce,  
an executive at Bell Labs  
who invented the word “transistor”  
and supervised development  
of the first communications satellite.



In 1966, John Pierce chaired the  
“Automatic Language Processing Advisory Committee” (ALPAC)  
which produced a report to the National Academy of Sciences,  
*Language and Machines: Computers in Translation and Linguistics*

And in 1969,  
he wrote a letter to the Journal of the Acoustical Society of America,  
“Whither Speech Recognition”

# The ALPAC Report

MT in 1966 was not very good, and ALPAC said diplomatically that

“The Committee cannot judge what the total annual expenditure for research and development toward improving translation should be. However, it should be spent hardheadedly toward important, realistic, and relatively short-range goals.”

In fact, U.S. MT funding went essentially to zero for more than 20 years.

The committee felt that science should precede engineering in such cases:

“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.”

John Pierce's views  
about automatic speech recognition  
were similar to his opinions about MT.

And his 1969 letter to JASA,  
expressing his personal opinion,  
was much less diplomatic  
than that 1966 N.A.S. committee report....

# “Whither Speech Recognition?”

“... a general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of a native speaker of English.”

“Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve ‘the problem.’ The basis for this is either individual inspiration (the ‘mad inventor’ source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach). . . .”

“The typical recognizer ... builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. **No simple, clear, sure knowledge is gained.** The work has been an experience, not an experiment.”

# Tell us what you really think, John

“We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn’t attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. **To sell suckers, one uses deceit and offers glamor.”**

“It is clear that glamor and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. Thus, we may pity workers whom we cannot respect.”

# Fallout from these blasts

The first idea: **Try Artificial Intelligence . . .**

DARPA Speech Understanding Research Project (1972-75)

Used classical AI to try to “understand what is being said  
with something of the facility of a native speaker”

DARPA SUR was viewed as a failure; funding was cut off after three years

The second idea: **Give Up.**

***1975-1986: No U.S. research funding for MT or ASR***



Pierce was far from the only person  
with a jaundiced view of R&D investment  
in the area of human language technology.

By the mid 1980s,  
many informed American research managers  
were equally skeptical about the prospects.

At the same time,  
many people believed that HLT was needed  
and in principle was feasible.

# 1985: Should DARPA restart HLT?

Charles Wayne -- DARPA program manager – has an idea.

He'll design a speech recognition research program that

- protects against “**glamour and deceit**”
  - because there is a well-defined, objective evaluation metric
  - applied by a neutral agent (NIST)
  - on shared data sets; and
- and ensures that “**simple, clear, sure knowledge is gained**”
  - because participants must reveal their methods
  - to the sponsor and to one another
  - at the time that the evaluation results are revealed

**In 1986 America,**

**no other sort of ASR program  
could have been gotten large-scale government funding.**

# NIST (Dave Pallett) 1985

"Performance Assessment of Automatic Speech Recognizers",  
*J. of Research of the National Bureau of Standards:*

Definitive tests to fully characterize automatic speech recognizer or system performance cannot be specified at present. However, it is possible to design and conduct performance assessment tests that make use of widely available speech data bases, use test procedures similar to those used by others, and that are well documented. These tests provide valuable benchmark data and informative, though limited, predictive power. **By contrast, tests that make use of speech data bases that are not made available to others and for which the test procedures and results are poorly documented provide little objective information on system performance.**

# “Common Task” structure

- A detailed “evaluation plan”
  - developed in consultation with researchers
  - and published as the first step in the project.
- Automatic evaluation software
  - written and maintained by NIST
  - and published at the start of the project.
- **Shared data:**
  - Training and “dev(elopment) test” data is published at start of project;
  - “eval(uation) test” data is withheld for periodic public evaluations

# Not everyone liked it

Many Piercians were skeptical:

“You can’t turn water into gasoline,  
no matter what you measure.”

Many researchers were disgruntled:

“It’s like being in first grade again --  
you’re told exactly what to do,  
and then you’re tested over and over .”

But it worked.

# Why did it work?

1. The obvious: it allowed funding to start  
*(because the project was glamour-and-deceit-proof)*  
and to continue  
*(because funders could measure progress over time)*
2. Less obvious: it allowed project-internal hill climbing
  - because the evaluation metrics were automatic
  - and the evaluation code was public*This obvious way of working was a new idea to many!  
... and researchers who had objected to be tested twice a year  
began testing themselves every hour...*
3. Even less obvious: it created a culture  
(because researchers shared methods and results  
on shared data with a common metric)

**Participation in this culture became so valuable  
that many research groups joined without funding**

# What else it did

The *common task method* created a positive feedback loop.

When everyone's program has to interpret the same ambiguous evidence, ambiguity resolution becomes a sort of gambling game, which rewards the use of statistical methods.

Given the nature of speech and language, statistical methods need the largest possible training set, which reinforces the value of shared data.

Iterated train-and-test cycles on this gambling game are addictive; they create “**simple, clear, sure knowledge**”, which motivates participation in the common-task culture.

# The past 25 years

Variants of this method

have been applied to many other problems:

*machine translation, speaker identification, language identification, parsing, sense disambiguation, information retrieval, information extraction, summarization, question answering, OCR, sentiment analysis, ... , etc.*

The general experience:

1. Error rates decline by a fixed percentage each year, to an asymptote depending on task and data quality
2. Progress usually comes from many small improvements; a change of 1% can be a reason to break out the champagne.
3. Shared data plays a crucial role – and is re-used in unexpected ways.
4. Glamour and deceit have been avoided.

**...and a self-sustaining process was started!**



## CoNLL (ACM SigNLL's annual meeting):

Since 1999, CoNLL has included a shared task in which training and test data is provided by the organizers which allows participating systems to be evaluated and compared in a systematic way. Descriptions of the participating systems and an evaluation of their performances are presented both at the conference and in the proceedings.

# The Linguistic Data Consortium

- Started in 1992 with seed money from DARPA
- LDC has distributed
  - more than 90,000 copies
  - of more than 1,300 titles
  - to more than 3,200 organizations
  - in more than 70 countries
- About half of the titles are “common task” datasets
  - developed for technology evaluation programs
  - released via general catalog after program use
- ~30 titles added to general catalog per year  
(current publications queue of about 120 items)

# Where we were

ANLP-1983

(First Conference on Applied Natural Language Processing)

34 Presentations:

None use a published data set.

None use a formal evaluation metric.

Two examples:

Wendy Lehnert and Steven Shwartz,

“EXPLORER: A Natural Language Processing System for Oil Exploration”.

Describes problem and system architecture; gives examples of queries and responses.

No way to evaluate performance or to compare to other systems/approaches.

Larry Reeker et al.,

“Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions”

Describes problem and system architecture; gives examples of inputs and outputs.

No way to evaluate performance or to compare to other systems/approaches.

# Where we are

ACL-2010

(48<sup>th</sup> Annual Meeting of the Association for Computational Linguistics)

274 presentations –

Essentially all use published data and published evaluation methods.  
(A few deal with new data-set creation and/or new evaluation metrics.)

Three examples:

Nils Reiter and Anette Frank, "Identifying Generic Noun Phrases".  
Authors are from Heidelberg University; use ACE-2 data.

Shih-Hsiang Lin and Berlin Chen,  
"A Risk Minimization Framework for Extractive Speech Summarization".  
Authors are from National Taiwan University;  
use Academia Sinica Broadcast News Corpus  
and the ROUGE metric (developed in DUC summarization track).

Laura Chiticariu et al., "An Algebraic Approach to Declarative Information Extraction".  
Authors are from IBM Research; use ACE NER metric, ACE data, ENRON corpus data.

# Social science is different...

But not that different:

Sharing data and problems

- lowers costs and barriers to entry
- creates intellectual communities
- speeds up replication and extension
- and guards against “glamour and deceit”  
(...as well as simple confusion)