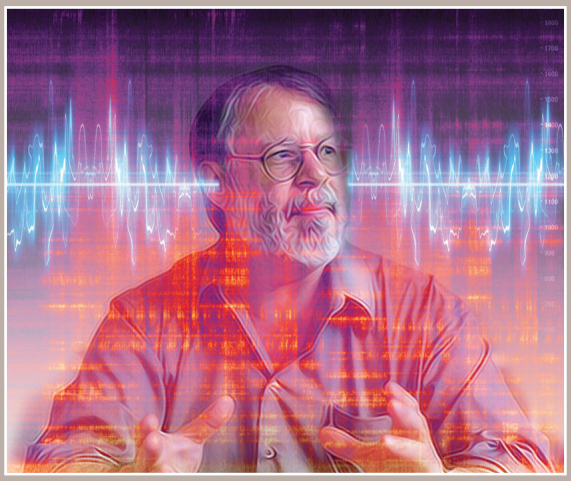
Mark Liberman

Why Speech Technology (almost) Works



Prisme N°32
December 2017

The Cournot Centre and Foundation

The Cournot Centre and Foundation

The Cournot Centre is an independent, non-profit organization. It has created a forum where specialists meet to advance theories and increase understanding of the economic and social sciences and their epistemology. Like the Centre, the Foundation pursues catalysing work in the tradition of Augustin Cournot, accelerating theoretical formulations. Under the aegis of the Fondation de France, the Foundation puts into perspective the probabilistic paradigm shift originating in mathematics and spreading across disciplines.

Why Speech Technology (almost) Works

Mark Liberman

Prisme N°32

December 2017

The Centre thanks Edouard Geoffrois for his careful proofreading.

Why does human language technology – speech recognition, speech synthesis, machine translation, information extraction from texts, question-answering, and so forth – almost work? What should scientists in other fields learn from the success of these disciplines? I would like to begin by reminding you of some of the ways in which human language technology works.

Let us take the example of the smartphone. Most phones now have some sort of speech-based question-answering input: there is Siri on iPhones, OK Google on Android phones and something similar on Windows phones. So this morning, I turned on my phone, and I said:

"OK Google, what is the French word for 'dog'?"

I figured it would probably transcribe that correctly, but I was surprised that not only did it transcribe that correctly, it answered via speech synthesis, "chien".

I thought that was pretty good. So then I asked:

"OK Google, what is 15 degrees centigrade in Fahrenheit?

It transcribed it correctly and answered right back:

"15 degrees Celsius is 59 degrees Fahrenheit".

That was answered in text, not in speech, but it could, of course, have synthesized it. Next I asked:

"OK Google, what's the name of the student newspaper at the University of Pennsylvania?"

It transcribed it correctly, answering with a page of search links with *The Daily Pennsylvanian* at the top, the student newspaper.

"OK Google, note to self: buy paper towels".

It transcribed:

"note to self: buy paper towels".

In this case, it sent me an email from myself saying that I should buy paper towels.

So now I figured I had to break it, do something it couldn't do. I happened to have an R graphics book called *ggplot2* by an author named Hadley Wickham. So I said:

> "OK Google, when was Hadley Wickham's book *ggplot2* published?"

I figured there was no way it was going to get that, and it did this weird thing, it transcribed:

> "When was Hadley Wickham's book *ggplot2* published?"

I have no idea how this book got into its lexicon. Then it showed me a page of search results with the Amazon listing for that book at the top. As I did not quite succeed in breaking it, I decided to ask something else that it probably would not know:

> "OK Google, what is the word for 'dog' in Hausa?"

Answer:

> "Here is your translation" [in speech synthesis]

Then it put me into Google translate with the word in Hausa. This was almost spooky! So I gave up trying to break OK Google, although I'm sure I could. If I took it into a noisy environment, for example, I'm sure that it would begin to fall apart.

Next I went to Google translate, and I cut and pasted something from the French website of the Cournot Centre:

> "Le Centre Cournot est une association soutenue par la Fondation Cournot, placée sous l'égide de la Fondation de France. Elle porte le nom du mathématicien et philosophe **franc-comtois** Augustin Cournot (1801–1877), reconnu de longue date comme un pionnier de la **discipline économique**."

Google translation:

> "The Cournot Centre is an association supported by the Cournot Foundation, under the aegis of the Fondation de France. It is named after the mathematician and philosopher

> **Franc-Comtois** Augustin Cournot (1801-1877), long recognized as a pioneer of **economic discipline**."

Here, there are a couple of mistakes. For example, Google translate does not know that "franc-comtois" is an adjectival form. It also says that Cournot is recognized as a pioneer of "economic discipline", which is not what "la discipline économique" means in French. "Economics" would be the correct translation, while "economic discipline" means something more like economic austerity measures.

Website:

> "Le Centre n'est pas un laboratoire de recherche, il n'est pas non plus un centre de réflexion. **Il** jouit de l'indépendance singulière d'un catalyseur."

Google translate:

> "The Centre is not a research laboratory, it is not a think tank. **He** enjoys the singular independence of a catalyst."

Notice in this segment that the pronoun "il", which in English should be "it" because it refers to the Cournot Centre, is translated as "**he** enjoys the singular independence of a catalyst...".

Website:

> "Pour qu'un débat ait lieu, il faut plus que de la connaissance et de la compréhension. Il faut des préférences, des croyances, des désirs, des objectifs… **C'est en pratique de cela seulement dont les débatteurs disposent et ils inventent ou ils adoptent les résultats qui leur conviennent**."

Google translate:

> "To have a debate, it takes more than knowledge and understanding. It takes preferences, beliefs, desires, goals ... **In practice this only with the debaters have and they invent or they adopt the results that suit them."**

In this final passage, Google translate does okay with the first two sentences, but the last becomes garbled, so at least Google translate occasionally messes up.

Finally, I have been reading a *roman policier* called, ***Le Dingue au Bistouri***,[1] and it starts this way:

> Il y a quatre choses que je déteste.
> Un: qu'on boive dans mon verre.
> Deux: qu'on se mouche dans un restaurant.
> Trois: qu'on me pose un lapin.
> […]

Google Translate:

> There are four things I hate.
> A: we drink in my glass.
> Two: we will fly in a restaurant.
> Three: I get asked a rabbit.
> […]

Google translate actually got the entire passage wrong. "Qu'on boive dans mon verre" should be that "someone drinks from my glass", but Google says, "we drink in my glass". "Qu'on se mouche dans un restaurant", has to do with blowing one's nose. Google translate confuses the verb "se moucher" and the noun "mouche" (meaning "fly"), producing the nonsensical, "we will fly in a restaurant". The phrase in the last line "pose un lapin" is an idiomatic expression meaning "to stand someone up". Google translate says, "I get asked a rabbit".

So, Google translate is not perfect, but you have to push a little bit to break it, at least for familiar languages like French and English. In the interest of fairness, I gave Bing translator a shot. It did a little worse if anything, but basically no better.

[1] Khadra, Yasmina (1983), Le dingue au bistouri, Paris: Éditions Poche.

Coming back to our main topic: today human language technology almost works. Why is that? What has happened? There are some obvious reasons. There is a kind of digital shadow universe that increasingly mirrors real life in flows and stores of bits. Society is mostly about communication, and most communication is text, or talk, which is just text in some sort of fancy writing, and more and more often in digital form. Simple properties of text (like the words that make it up) are a good proxy for content. We have bigger, faster, cheaper digital everything – networking, computers, phones,… – and better programing languages, and so on, that make it easier to pull content out of the flows of text in this digital shadow universe.

There is an old argument about whether "content is king" or "communication is king". But the "content of communication" is at least the power behind the throne. So, in this new evolutionary niche, new life forms – such as OK Google and Siri, and so on – have the means, motive and opportunity to live in that ecological niche, while adding their own digestion products to the ecosystem. That is one reason why human language technology is getting to be quite good: it creates an opportunity.

Another reason that it almost works is obviously advances in machine learning, which is basically applied statistics and the computer power to apply them. There are all sorts of acronyms and buzzwords, such as "long short-term memory" and "deep neural nets", and so on, that represent interesting, often conceptually simple, but mathematically complex techniques that can be applied to solve problems in speech and language analysis, and an enormous amount of progress has been made in the last couple of decades along those lines.

But there is another reason, which I believe is more important than either of the two that I just gave. It is a cultural change that took place half a century ago, and the rest of this text is about that story.

What I am going to tell you is based on a presentation that I gave in 2015 during a workshop at the National Academy of Sciences on, "Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results".[2] The workshop was

[2] Michelle Schwalbe, Rapporteur; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Academies of Sciences,

alarming, and it was meant to be alarming, because there is a crisis of credibility in many areas of scientific research, as widely documented. There is a famous paper by John Ioannidis, "Why Most Published Research Findings are False"[3], which was published in 2005 and has stood up quite well during the intervening time. More recently in the *Chronicle of Higher Education*, there was an article entitled, "Amid a Sea of False Findings, the NIH Tries Reform".[4] They quoted Dr. Francis Collins, the director of the National Institutes of Health in the United States, saying that Amyotrophic Lateral Sclerosis (ALS) researchers seeking a cure for that disease went back and tried to reproduce studies on more than 70 promising drugs. "Zero of those were replicable," Dr. [Francis] Collins said. "Zero. And a couple of them had already moved into human clinical trials …". So people around the world are concerned about this. I learned that the psycholinguists at the École Normale Supérieure (ENS) have been holding a series of workshops on reproducibility in psychology, but it is not just in psychology, it is also in biomedical research and many other fields. So I am going to tell the story of a crisis of credibility that afflicted a different research area some 50 years ago.

Once upon a time, there was a Bell Labs executive named John Pierce. He supervised the team that built the first transistor; he oversaw development of the first communication satellite. He had no problem with credibility. The photograph below is of him; that is what engineers looked like in the 1950s: they wore suits and ties, and they sat next to complicated analog equipment with lots of dials and switches. In 1966, Pierce chaired a committee – the Automatic Language Processing Advisory Committee, known familiarly as ALPAC – which produced a report to the National Academy of Sciences on machine translation. The ALPAC report[5] noted that machine

---

Engineering, and Medicine (2016), Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop, Washington, D.C.: The National Academies Press.

[3] Ioannidis, John (2005), "Why most published research findings are false", PLoS Medicine, 2(8), August, e124.

[4] Voosen, Paul (2015), "Amid a Sea of False Findings, the NIH Tries Reform", The Chronicle of Higher Education, 16 March.

[5] ALPAC (1966), *Language and Machines: Computers in Translation and Linguistics.* A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council, Publication 1416.

translation in 1966 was not very good, and suggested the following in "diplomatic" (in the language of committees) terms: "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." (ALPAC, 1966, p. 33). That is committee language for "stop giving them money!"

Image 1: Photograph of John Pierce, circa 1950s

The Committee felt that science should precede engineering in such cases, and they more or less pointed in the direction of the wonderful possibilities that computers offered to linguistic research. The funders, however, read between the lines, and machine translation funding in the United States went to $0 for more than 20 years after that report was issued. Pierce's views about automatic speech recognition were similar to his opinions about machine translation. In 1969, he wrote a letter to *The Journal of the Acoustical Society of America* published under the title "Whither Speech

Recognition?",[6] in which he expressed his personal opinion, phrased in less diplomatic language. He wrote:

> *...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* [and by that he means researchers working on recognition] *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.*

He then went on to say:

> *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%.* [This is probably false from the point of view of empirical economics: if you could cut the cost of soap by 10 per cent, you could

[6] Pierce, John (1969), "Whither Speech Recognition?", *The Journal of the Acoustical Society of America,* 46(4), 1049–1051(L).

> attract investors…] *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.* (Pierce, 1966, pp. 1049-1051)

That is strong language. Then various luminaries in the nascent field of artificial intelligence argued back that the whole problem was that those "mad inventors" and "untrustworthy engineers" did not understand things like Lisp (List processing computer language), artificial intelligence, and first-order predicate calculus applied to problems of pattern recognition. So they persuaded the Defense Advanced Research Projects Agency (DARPA)[7] to invest in a program that would try to apply artificial intelligence to the problem of speech recognition. The program tried to use classical AI – applied logic – to understand what is being said with something of the facility of a native speaker. It leaned heavily on *a priori* ideas of what was being said.

One of the programs that they built was for playing chess with a robot by saying your moves rather than doing something on a chessboard or with a graphical user interface. You could say, "pawn to queen two", or something like that, and it would move the pawn if it understood you. At one of the demonstrations of this project, one of the skeptical DARPA officials who had read John Pierce's screed played a game of chess with this machine by barking at it like a dog. He would bark "woof, woof", and it would make some plausible move, because it had an idea of what the possible semantic space was, and it would pick a move that would best fit the acoustics it was given. So this project was viewed as a failure and funding was cut off somewhat prematurely after three years. The second idea was to give up. Between 1975 and 1986, there was simply no significant research funding in the United States for machine translation or automatic speech recognition.

[7] DARPA is an agency of the US Department of Defense responsible for the development of emerging technologies for use by the military.

John Pierce was not the only person who had a negative view of R&D investment in this general area, and many, maybe most informed American research managers, were equally skeptical about the prospects. When I began working at Bell Labs in 1975, only the most limited, tentative, small isolated-digit, or maybe connected-digit, kind of recognition projects could even be countenanced, and those were viewed as not likely to be successful. Nevertheless, at the same time, there were many people who believed that human language technology was needed, or would, in the future, be needed and, in principle, ought to be feasible. By 1985, this came to a head, and there was a big debate within certain circles about whether or not DARPA should restart human technology research.

I should mention that DARPA is an unusual government agency, not only for the United States but perhaps for the world, in that it has a very large budget, but almost no permanent employees. All of its program managers and their staff are rotated in from universities or from other government jobs, and the administrative staff is contracted out to local groups who supply people on a limited-term basis for a given program that is intended to last for a limited period of time and then disappear.

Charles Wayne, who was "on loan" to DARPA from the National Security Agency, had an idea. He came to DARPA to head a human language technology program in 1985 and had to overcome considerable skepticism and negative reaction on the part of his management. So he decided that his program would protect against "glamour and deceit", because there would be a well-defined objective, quantitative, evaluation metric that would be applied by a neutral third party – the National Institute of Standards and Technology (NIST) – on shared datasets, unpublished data. It would ensure that simple, clear, sure knowledge was gained, because the participants would be required to reveal their methods to the sponsor and to one another at the time that the evaluation results were presented.

In order to do this, he needed published data, of course with some withheld for testing, and well-defined metrics. He thus enlisted David Pallett at what was then called the National Bureau of Standards and is now the National Institute of Standards and Technology. Pallet looked into the matter and wrote a paper for the

Journal of Research of the National Bureau of Standards[8] in 1985 on performance assessment of automatic speech recognizers. He wrote:

> *Definitive tests to* fully *characterize automatic speech recognizer or system performance cannot be specified at present* [by which he meant that there are many kinds of listening conditions, speech, speakers, topics, and so on.]. *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.*** [author's emphasis] (Pallett, 1985, p. 371)

Shortly before this article was published, George Doddington, an electrical engineer working for Texas Instruments at the time, had done something that caught the eye of people at DARPA, including Charles Wayne, and people at the NIST including David Pallett. When he went to work for Texas Instruments, the first thing he did was to produce a small, cheap LPC[9] synthesis chip that was used in a toy called the *Speak & Spell*, which was an approximately $50 digital electronic toy. It was probably the first serious digital electronic toy. You would press the button, and it would say, "spell cat", and you would press "c-a-t", and it would tell you if you were right or wrong, and so on. Because in those days memory was very expensive – you could not store audio waveforms for hundreds or thousands of words – they used LPC synthesis with extremely low bandwidth compressed speech. The Speak & Spell was a big success, and he made a lot of money for Texas Instruments. When they asked him

[8] Pallett, David (1985), "Performance Assessment of Automatic Speech Recognizers", *Journal of Research of the National Bureau of Standards,* 90(5), September-October, pp. 371-387.

[9] Linear predictive coding.

what he wanted to do next, he said that he wanted to work on speech recognition. The first thing he did was to go out and buy one each of all of the programs or devices that were then sold. They mostly came from Japan at that time, because companies there had continued to work on this problem. They, of course, promised 97.6 per cent accuracy, and so on. Doddington then created his own database to test them, which was a database of connected digits spoken in American English by a large number of speakers. He tested each of them on this database of connected digits and worked out how well they did by some straightforward metric that he had devised. He wrote a paper[10] about this, which he submitted to the *IEEE Spectrum* – a glossy, broad circulation magazine sent to all the approximately 30-40,000 IEEE[11] members.

Because he worked for a company, he had to submit the article for publications clearance, which is often common for companies. They immediately rejected it on the grounds that this was company proprietary information. The company had spent a lot of money determining the level of the state-of-the-art of the specific devices and how well they worked, so they were against the idea of letting that information out into the industry. Doddington went ahead and published it anyway. The magazine even made it the cover story. So the then vice-president of research at Texas Instruments, Richard Wiggins, of course, received his copy in the mail and saw this cover story. He immediately called Doddington into his office, and red in the face, waved the magazine at him and said, "Doddington, what in the hell do you call this?" George sat down, smiled and said, "Well, I guess I would call it blatant insubordination. The question is: what are you going to do about it?" What they did about it, of course, was that they promoted him and gave him some larger projects to work on. This example shows that the idea of doing something like that in order to get a plausible benchmark evaluation on a limited task of a large number of competing recognizers seized the imagination of some of the people involved in funding such research.

---

[10] Doddington, George, and T.B. Schalk (1981), "Computers: Speech recognition: Turning theory to practice: New ICs have brought the requisite computer power to speech technology; an evaluation of equipment shows where it stands today", *IEEE Spectrum,* 18(9), September, pp. 26-32.

[11] Institute of Electrical and Electronics Engineers

The result of this was something that came to be called the "common task structure", which began with a detailed task definition and evaluation plan that was developed through a sometimes lengthy process of iterative consultation with researchers. The program manager would say what he wanted, and the researchers would come back to him saying they could not possibly do that; they would suggest they try something else, and so on. They went back and forth until they agreed on something. This was then published as the very first step in the project when they asked people for bids in doing research. Next, they hired the NIST to develop automatic evaluation software, which was also published at the start of the project. They commissioned the creation of training and development test (or "dev test") data, which was also given out at the start of the project; and they would withhold the evaluation data for periodic public evaluations.

Not everybody liked this. There were a lot of people, including at the time John Pierce, who were skeptical. Their attitude was: "you can't turn water into gasoline, no matter what you measure". Researchers, on the other hand, were very upset. Richard Schwartz at Bolt Beranek and Newman (BBN) – who had been one of the prominent people in the 1972–1975 DARPA speech understanding research projects, and for many years a leader of BBN's very important and internationally competitive speech recognition efforts – said, "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. It worked for one obvious reason: it allowed the money to flow; it allowed them to start paying people to work on the problem. Some kinds of problems get solved by accident while somebody is trying to do something else, but speech recognition is not likely to be one of those problems. It also allowed funding to continue, because the funders could measure progress over time. That was very important, because they started in 1985–1986, and it took 25 years or more – until very recently – to have anything resembling commercial success, any products, things the military could use or that people would pay to get, that used these technologies.

A less obvious reason that it worked was that it allowed project internal hill climbing, because the evaluation metrics were automatic and the evaluation code was public. So an obvious way of working emerged, which came as a revelation to many of the researchers. The same researchers who had objected to being tested twice a

year began testing themselves over and over again as fast as they could rewrite their code: every hour, every day or every week.

Perhaps an even less obvious reason that it worked was that it created a culture, because the 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. An early example is one of the first text retrieval conferences (TREC), which was funded by the US Defense department. They had funded four "performers" or sites that had contracts to do research in the area, but 40 laboratories signed up to participate in the evaluation and to talk about the research they had done. They realized that, just by creating the shared data, the published evaluation specification and the evaluation framework, they could get research done in an area without paying anyone to do it.

It also changed the nature of pattern recognition research, in general, and speech and natural language processing research, in particular. When everybody's program has to interpret the same ambiguous evidence, ambiguity resolution becomes a gambling game. This rewards the use of statistical and probabilistic methods, which led directly to the flowering of machine learning. So artificial intelligence, which in the 1970s and 1980s had been applied logic, became applied statistics for the most part. A little bit of logic is starting to creep back in slowly as the new generation comes along, but it really changed things in a major way.

Given the nature of speech and language, statistical methods also need the largest possible training set, which reinforces the value of shared data since groups can generally afford larger bodies of material than individuals can obtain. The iterated train-and-test cycle on this gambling game are, I think, literally addictive (I suspect that there are dopamine releases involved in this). In addition to making evaluation addicts out of researchers, they create simple, clear, sure knowledge, which motivates continued participation in the common task culture. They become like the people in the gambling halls who put their money into the machine and pull the levers, except that they're doing something more productive.

So the common task method has become the standard research paradigm in experimental computational science, not just in speech and language technology. This involves published training and testing data, well-defined evaluation metrics, various

kinds of techniques to avoid overfitting, which include very scrupulous avoidance of testing on the training data, but also withholding some evaluation data for a genuinely independent evaluation. These are managerial as well as statistical methods. The domain concerned by this common task method includes about anything – any kind of algorithmic analysis of the natural world. There may be other areas as well, but for trying to interpret facts, raw observations about the world, and so on, this is the main domain.

Since 1985, variants of this method have been applied to at least dozens of other problems: machine translation, speaker and language identification, parsing, sense disambiguation, information retrieval and extraction, summarization, question-answering, optical character recognition, sentiment analysis, image analysis, video analysis, and so on, and even autonomous vehicle navigation and certain areas of robotics. The general experience is that error rates declined by a fixed percentage every year, that is, performance gets exponentially better to an asymptote that depends on the task and on the quality of the training and testing data.

Interestingly, progress usually comes by many small improvements. It can be a little depressing. You go to a conference or a workshop in this area, and there are dozens of presentations, and each of them describes some deep conceptual analysis, some complicated mathematics, some difficult programing, weeks of computer time on very fast computers..., and it improves the performance on a standard benchmark by a third of a percent, and that is viewed as reason to pop open the Champagne! Why the celebration? It's because several thirds of a percent improvement add up to something. In fact, if you can improve things by 1 percent, then that is really great. When people using so-called deep neural net methods managed to improve performance on one standard speech recognition benchmark in a way that cut error rates by one-third from whatever it was – say from 20 percent to 14 percent – that was amazing, the biggest single change that had happened in decades. Shared data plays a crucial role, and it is often re-used in unexpected ways. Glamour and deceit have mostly been avoided. As commercial success emerges, of course, the temptation grows, and we're beginning to see a bit of it returning, but relatively little.

There are dozens of current examples. Some of them are shared task workshops, such as a series of annual workshops under the acronym of CoNLL, the

conference on natural language learning. There is open keyword search evaluation (OpenKWS) and open machine translation evaluation (OpenMT). In France, there is REPERE: REconnaissance de PERsonnes dans des Emissions audiovisuelles. Others include: Speaker Recognition Evaluation, Text Retrieval Conference (which occurs periodically), Shared Task on Pronoun Translation, TREC Video Retrieval, IMAGENET Large Scale Visual Recognition, and many, many others. Every week I get a few emails telling me about some that I have not heard of before. Some are just shared datasets and evaluation metrics, such as the Text Analysis Conference (TAC). It does not pay anybody to do any research; it is a series of evaluation workshops encouraging research in natural language processing (NLP) and related applications, by providing a large test collection, common evaluation procedures and a forum for organizations to share their results. TAC comprises sets of tasks known as "tracks", each of which focuses on a particular sub-problem of NLP. TAC tracks focus on end-user tasks, but also include component evaluations situated within the context of end-user tasks. A recent call was on knowledge-based population and biomedical summarization. TRECVID promotes content analysis and retrieval from digital video, so they are doing semantic indexing, interactive surveillance event detection, instance search in a BBC sitcom, multimedia event detection, localization, and video hyperlinking.

The Google Street View house numbers dataset is a very interesting recent case. They took house numbers automatically recognized in Google Street View images, and they published a set of more than 73,000 digits with human identification of the digits, 26,000 digits where they withheld the information for testing, and another set of somewhat more than half a million samples for extra training. The progress in performance was unusually rapid. In 2011, the error rate was 36.7%; in 2015, it was 1.92%. Progress is not always this rapid, but steady progress almost always happens when this method is used. Here is what a sample of them looks like.

Image 2: Sample of a Google Street View house numbers dataset

The next figure represents a famous graph that shows speech-to-text benchmark test history through 2009. The point is that in most cases the curves are trending downward. I remember when they first tried the switchboard task back in the early 1990s, the error rate was something like 90 per cent. Rather than being discouraged, people thought that was great, because it meant that the program would continue for many years. If you try something and you get a 10 per cent error rate, you know 5 per cent is the noise floor, so you probably do not have very many years of funding in that project. There is some continued progress in speech-to-text, so the switchboard Corpus had sort of stalled at about a 20-to-30 per-cent error rate 15 years ago, but it is down to about 10 per cent now, bringing it within hailing distance of typical human disagreements about what people actually said, which are in the range of 5 to 6 per cent.

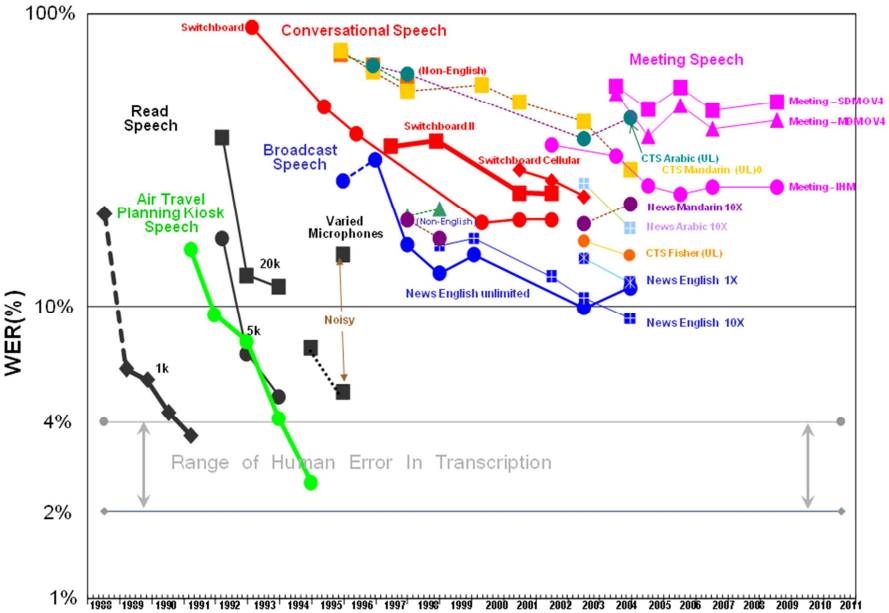


Figure 1: NIST Speech-to-Text Benchmark Test History, May 2009

In 1983, the first conference on applied natural language rocessing had 34 presentations, not a single one of which used a published dataset or a formal evaluation metric. In 2010, at the 48th annual meeting of the Association for Computational Linguistics, there were 274 presentations, and every single one used published data and published evaluation methods, with the exception of perhaps three that described new dataset creation or a new evaluation metric creation.

I hope that this text conveys the intellectual history of what is now a completely firmly ingrained, established culture among people who work on human language technologies in general. Newer researchers in this area are sometimes not aware that anyone ever did it differently. But what is the intellectual history of other areas, and, in particular, outside of engineering, at least what we call science? Many areas of science are not that different, because sharing data and problems lowers costs and barriers to entry. Let us take the example of something that has been of great interest in the United States recently and worldwide: projecting the development of neurodegenerative disorders in older people, especially Alzheimer's

disease. Imagine I go to the doctor and say, "I'm having more trouble remembering telephone numbers and names than I used to, and I wonder whether something is happening". The doctor will give me a test. Let's say he discovers that my digit span is smaller, and maybe on the low side for even someone my age. I then ask, "What's going to happen to me? What's the future going to bring?" He can take blood, take pictures of my brain with magnetic resonance imagery, take cerebrospinal fluid, do a gene scan, and in the end, he cannot tell me anything except that in 10 years maybe I'll be a vegetable and maybe I'll be just like I am now, only a little more forgetful.

In a way, that's a lot like speech recognition, that is, you have potentially a lot of data, data about how people talk, how they write, what their digit span is – there are other kinds of psychological tests –, about blood work, brain scans, genotype, and so on. Maybe from that, if you get longitudinal data about lots of people like that, somebody ought to be able to figure out how to predict something. But at the moment you could not even try that unless you were involved in a clinical group that had lots and lots of patients of that kind. No single group has very many patients of that kind, maybe a few hundred. So in 2004, the NIST organized the "Alzheimer's Disease Neuroimaging Initiative" (ADNI), which involved 30 clinical sites. Neil Buckholtz of the National Institute on Aging was involved in starting this. We were on the same panel, "Transforming Research through Open Access to Discovery Inputs and Outputs" at a conference in Berlin in 2011, and he gave a talk about the ADNI that surprised me enormously. One of his slides including the following goals of the ADNI longitudinal multi-site observational study:

- Collect data and samples to establish a brain imaging, biomarker, and clinical database in order to identify the best markers for following disease progression and monitoring treatment response.
- Determine the optimum methods for acquiring, processing and distributing images and biomarkers in conjunction with clinical and neuropsychological data in a multi-site context.
- "Validate" imaging and biomarker data by correlating with neuropsychological and clinical data.
- Provide rapid public access of *all* data and access to samples.

If you want this data, you cannot just download it over the Internet. You have to go to their website and send them a letter saying who you are, why you want it, and what you propose to do with it. If you do that, they'll give it to you. And if you have a bright idea about how to do better than other people at predicting the progress of neurodegenerative disorders, you will probably get the Nobel Prize in medicine. Something, however, is lacking in this ADNI: there is no well-defined versioning of the datasets. You just get whatever they've accumulated up to the time that you sign the agreements and they send you the data. There is no evaluation metric, so there is no way to compare your results against other people's results. You just get the data, because obviously biomedical researchers know how to do research; you don't have to give them an evaluation metric. There are no focused workshops in which people compare their results on this data. They just publish work in the literature if they feel like it. In my opinion, predicting the time course of Alzheimer's disease is exactly the kind of problem for which the common task method seems to work. It seems to me that we ought to consider applying such methods to the rather large class of similar biomedical problems. I think initially scientists would mostly be horrified; they would have the same reaction that Richard Schwartz did – that the bureaucrats are taking away their investigators' freedom to pursue things the way they want and making them all act in lockstep. But maybe we should consider it.

The *Prisme* Series

The *Prisme* Series is a collection of original texts that focus on contemporary theoretical issues. The authors are contributors to the Cournot series of conferences, panels and seminars.

Latest releases:

35. Big Data: A Game Changer for Quantitative Finance?
Mathieu Rosenbaum

34. Saturation and Growth: When Demand for Minerals Peaks
Raimund Bleischwitz and Victor Nechifor

33. A Time-Frequency Analysis of Oil Price Data
Josselin Garnier and Knut Sølna

31. How to Flee on a Straight Line: Tracking Self-Repulsive Random Walks
Laure Dumaz

30. The Evolving Connection between Probability and Statistics
Noureddine El Karoui

A complete list of publications can be found at
www.centre-cournot.org

Mark Liberman is a Professor of Phonetics, Linguistics and Computer and Information Science at the University of Pennsylvania and the Founder and Director of the Linguistic Data Consortium. He obtained his Ph.D. in linguistics from the Massachusetts Institute of Technology in 1975 and began his career at AT&T Bell Laboratories that same year. He was a member of its technical staff from 1975 to 1987 and the Head of its Linguistics Research Department from 1987 to 1990. Liberman is the author of three books and over 100 published papers.

His research interests include corpus-based phonetics; speech and language technology; the phonology and phonetics of lexical tone, and its relationship to intonation; gestural, prosodic, morphological and syntactic ways of marking focus, and their use in discourse; formal models for linguistic annotation; information retrieval and information extraction from text; the organization of spoken communication in the human brain, especially in relation to the evolutionary substrates for speech and language, and to analogous systems in other animals; agent-based models of language evolution and learning.

Series Editor: Jean-Philippe Touffut
Illustration Artist: Gerald Wassen / Cover Designer: Sophie Otrage