Obituary

Fred Jelinek

Mark Liberman
University of Pennsylvania

Frederick Jelinek died, peacefully and unexpectedly, on 14 September 2010. Over a distinguished career of nearly fifty years, Fred made important contributions in areas ranging from coding theory and speech recognition to parsing and machine translation. But more important than these specific contributions was his role in persuading the fields of speech and language engineering to adopt statistical methods and the "noisy channel model," returning to the path opened up by Claude Shannon in 1948. And more important still was the role that he played in defining, exemplifying, and promoting what has become the standard research paradigm in an even wider range of fields: the competitive evaluation of alternative algorithms based on precise quantitative criteria defined in advance, relative to a shared body of training and testing material.

After receiving his Ph.D. from MIT in 1962, Fred taught at Cornell from 1962 to 1972, worked at IBM from 1972 to 1993, and taught at Johns Hopkins from 1993 to 2010. Fred's many technical accomplishments during this long and productive career can be seen as episodes in two epic narratives, which, like the *Iliad* and the *Odyssey*, are related but have separate themes and story lines. The theme of the first epic is the return of Information Theory to center stage in speech and language processing; and the theme of the second epic is the development of a new relationship between science and engineering in speech recognition, computational linguistics, and artificial intelligence (AI) more generally.

Fred gave us a vivid first-person narrative of the first of these epic adventures in his ACL Lifetime Achievement Award speech (Jelinek 2009). But missing from this recital is a picture of the dire state of the information-theoretic forces when Fred joined the fray by undertaking speech recognition research at IBM in 1972. To understand what the world was like then, we need to go back to the mid 1950s.

Claude Shannon (1956, page 3) wrote:

> Information theory has, in the last few years, become something of a scientific bandwagon. Starting as a technical tool for the communication engineer, it has received an extraordinary amount of publicity in the popular as well as the scientific press. . . . Our fellow scientists in many different fields, attracted by the fanfare and by the new avenues opened to scientific analysis, are using these ideas in their own problems. Applications are being made to biology, psychology, linguistics, fundamental physics, economics, the theory of organization, and many others. . . .
>
> Although this wave of popularity is certainly pleasant and exciting for those of us working in the field, it carries at the same time an element of danger. . . . It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like information, entropy, redundancy, do not solve all our problems.

Ironically, the seeds of that overnight collapse were at that very moment being sown by a young linguist with a fresh Ph.D. from Penn, whose lecture notes for an MIT

undergraduate course were being prepared for publication in the Netherlands. Noam Chomsky's (1957, page 16) opus *Syntactic Structure*s famously argued that

> It is fair to assume that neither [(1) "Colorless green ideas sleep furiously"] nor [(2) "Furiously sleep ideas green colorless"] (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. Yet (1), though nonsensical, is grammatical, while (2) is not.

Although this assertion was not further justified—and would later be shown to be false as a general indictment of "any statistical model for grammaticalness" (e.g., in Pereira 2000)—it plausibly and effectively undermined the value of empirical *n*-gram approximations as models of natural-language syntax. And more importantly, *Syntactic Structures* pointed the way to larger issues in mathematical and computational linguistics, whose exploration occupied many productive years of research over the next several decades. At the same time, AI researchers were promoting the view that intelligence should be seen as a matter of applied logic, with everything from perceptual classification and motor planning to chess playing and conversational interaction being modeled as particular kinds of theorem proving.

As a result, by the time that Fred Jelinek went to IBM in 1972 to work on speech recognition, the Information Theory bandwagon of the 1950s was lying forgotten in a ditch, at least as far as applications in speech and language processing were concerned. One personal memory may help to make this point to those who didn't live through the period.

In the 1980s, Ken Church was one of the heroes fighting effectively for corpus-trained statistical methods in computational linguistics. But in the late 1970s, when I first met him, Ken embodied a very different set of attitudes. He was working on an MIT dissertation applying context-free parsing to phonotactic speech recognition, assuming a non-stochastic grammar. I was curious about the redundancy of phonotactics in general, and also felt that a stochastic phonotactic grammar would do a better job, so I suggested that Ken should find or create a collection of phonetic transcriptions, and use it to associate probabilities with his rewrite rules. Ken's response was to quote Richard Nixon's remark about Daniel Ellsberg: "We could kill him—but that would be wrong." Further discussion elicited a quote from one of his AI lab professors: "If you need to count higher than one, you've made a mistake."

The partisans of what John Bridle has called the "cybernetic underground," with Fred Jelinek as one of their most important leaders, soon taught Ken and others of his generation why it is a good idea to count higher than one, to use the counts to estimate the parameters of statistical models, and to use those models to resolve ambiguity. Fred's success in this struggle was tied to the second epic adventure in which he played a role. And again, the stage for this second story was set while he was working on information theory at Cornell.

It started in the 1960s with two interventions by John R. Pierce, then a high-ranking executive in the research area at Bell Laboratories. The effect of these interventions was to halt most research in the United States in the areas of machine translation and speech recognition, on the grounds that the scientific foundations of these fields were not adequately established, and that engineering work in advance of fundamental science was a waste of time and money. This general view of the relationship between science and technology was typical of the men who led U.S. engineering research during World War II and the post-war period, and John Pierce had both the credentials and the

confidence to apply these ideas in an authoritative way, and to persuade funders and researchers to his point of view. Pierce had supervised the team that developed the transistor, and coined the word *transistor* itself; he supervised the development of the first communication satellites; and he had made contributions in areas as far afield as computer music and science fiction.

The topic of John Pierce's first intervention was machine translation. He chaired a committee of the National Academy of Sciences, whose 1966 report (Pierce and Carroll 1966, page 30) persuaded the U.S. government to stop funding machine translation (MT) research. It is well known that the ALPAC report torpedoed MT funding in the U.S., but it is less well known that it promoted fundamental research in computational linguistics as an alternative:

> 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.

Pierce's second intervention came in the form of a letter to the *Journal of the Acoustical Society of America* (Pierce 1969). Because it was not written by a committee, it was much more direct:

> 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 glamour.
>
> It is clear that glamour 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 (page 1049).

The key problem, Pierce thought, was failure to build on past accomplishments in the way that successful science and engineering do:

> 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 (page 1050).

Pierce argues that "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" (page 1050). The clear implication is that engineers should stop trying to solve the general speech recognition problem until the relevant scientific foundations have been laid.

This letter had an even more devastating effect than the ALPAC report did. While Fred was leading IBM's effort to solve the general dictation problem during the decade or so following 1972, most other U.S. companies and academic researchers were working on very limited problems such as speaker-dependent small-vocabulary isolated-word recognition, or were staying out of the field entirely.

However, Fred's information-theoretic approach provided a purely engineering-based way to make speech recognition research an "experiment" rather than an "experience." Comparing the quantitative performance of alternative algorithms on stable training and testing sets, using fixed and automatically calculated evaluation metrics, offered a nearly glamour- and deceit-proof way out of Pierce's trap.

This was the mode of work in Fred's group at IBM, first applied to speech recognition and later to a variety of other problems, including especially machine translation. From the beginning, Fred was strongly in favor of sharing resources, evaluation metrics, and algorithmic ideas with researchers outside of IBM. In fact, some of the seeds of DARPA's efforts in this area developed out of Fred's efforts, beginning in the mid 1980s, to decrease his group's costs by sharing the development of expensive resources such as dictionaries and parallel text corpora. Under Fred's leadership, in 1989 IBM donated its aligned version of the Canadian Hansards—the parallel text corpus used in his pioneering MT research—to the ACL's Data Collection Initiative.

Charles Wayne, starting a new speech-recognition program at DARPA in 1986, adopted the idea of quantitative comparison of alternative algorithms on a fixed task. In the context of this program, the formal quantitiative competition was not only among algorithms but among research groups, with all the sites involved in the project sharing a predefined automatic evaluation metric and a body of material for training and testing.

At first, DARPA's key motivation for this "common task method" seems mainly to have been to persuade skeptics that the new program would successfully avoid glamour and deceit. And at first, many researchers complained that the frequent quantitative evaluations made them feel as if they had returned to kindergarten. But it soon became clear that this approach had a number of key advantages as a method of research management. It lowered barriers to entry, by providing well-defined tasks and the resources needed to undertake them; it created a research community with shared goals and assumptions; and, perhaps most important, it offered proof of gradual progress that could be used to justify stable funding for speech and language technology development over several decades, even when no "killer app" had yet emerged.

As a result of this demonstrated success, DARPA's investment in Human Language Technology continued, and grew to include text retrieval, information extraction, topic detection and tracking, summarization, machine translation, and many other problems. And speech and language researchers have come to take this paradigm for granted, and to apply it even when it is not imposed by funders.

Contrary to the expectations of many well-informed people, this approach has allowed a very large number of small algorithmic improvements to accumulate over several decades, to the point where without any major breakthroughs, speech recognition and machine translation are now quite workable in many applications.

And this history now puts us in a position to apply John Pierce's perspective back-to-front. Like most of the rest of his generation, Pierce felt that the natural progression was for scientific understanding to serve as the foundation for engineering applications. But instead, more than three decades of engineering development in human language technology may now permit rapid scientific progress.

Independent of their value in practical applications, the algorithms developed by the process that Fred Jelinek pioneered offer marvelous new tools for scientists. Applying these tools to the vast stores of digital speech and text now becoming available, we can observe linguistic patterns in space, time, and cultural context, on a scale many orders of magnitude greater than in the past, and simultaneously in much greater detail. Rather than evoking the impact of particle accelerators, as the ALPAC report did, it may be more appropriate to compare these tools to the invention of the microscope and telescope in the 17th century: Everywhere we look, there are interesting patterns previously unseen.

In the published version of his 2009 ACL Lifetime Achievement Award speech, Fred Jelinek wrote (page 484):

> I sat in [on Noam Chomsky's lectures in 1961] and got the crazy notion that I should switch from Information Theory to Linguistics. I went so far as to explore this notion with Professor Chomsky. Of course, word got around to my adviser Fano, whom it really upset. He declared that I could contemplate switching only after I had received my doctorate in Information Theory. I had no choice other than to obey. Soon thereafter, my thesis almost finished, I started interviewing at universities for a faculty position. After my job talk at Cornell I was approached by the eminent linguist Charles Hockett, who said that he hoped that I would accept the Cornell offer and help develop his ideas on how to apply Information Theory to Linguistics. That decided me. Surprisingly, when I took up my post in the fall of 1962, there was no sign of Hockett. After several months I summoned my courage and went to ask him when he wanted to start working with me. He answered that he was no longer interested, that he now concentrated on composing operas.
>
> Discouraged a second time, I devoted the next ten years to Information Theory.

Perhaps, in the end, this was the best way for Fred to contribute to linguistics. He lived to see the triumph of his ideas in speech and language engineering; we should remember him as we explore the world that his ideas are opening up to us in speech and language science.

References

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague.

Jelinek, Frederick. 2009. The dawn of statistical ASR and MT. *Computational Linguistics*, 35(4):483–494.

Pereira, Fernando. 2000. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society A*, 358(1769):1239–1253.

Pierce, John R. 1969. Whither speech recognition? *Journal of Acoustical Society of America*, 46(4B):1049–1051.

Pierce, John R. and John B. Carroll. 1966. *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences/National Research Council, Washington, DC.

Shannon, Claude. 1956. The bandwagon. *IRE Transactions on Information Theory*, 2(1):3.