TECHNOCULTURE

Speech and language technology: People are not dictionaries

Episode 8

Full transcript

Guest: Mark Liberman [Mark]
Host: Federica Bressan [Federica]

[Federica]: Welcome to a new episode of Technoculture. I'm your host, Federica Bressan, and today my guest is Mark Liberman, an American linguist who has a dual appointment at the University of Pennsylvania, both at the Department of Linguistics and at the Department of Computer and Information Sciences. Mark is also founder and director of the LDC, the Linguistic Data Consortium. Welcome, Mark.

[Mark]: Glad to be here.

[Federica]: So what fascinates me about computational linguistics is that it's a field where two different disciplines — computer science and linguistics — have encountered each other quite a long time ago. That means that today, computational linguistics has some history. We can see how these two disciplines merged and now operate together, and that fascinates me because my current research is placed in the digital humanities, which is not technically such a new thing, but it's recent, and people still debate how to, in fact, put together these different disciplines and see them operate together with a scientifically credible methodology.

So I would like to ask you, to begin with: How does this marriage between linguistics and computer science work, and when did it start?

[Mark]: You know, I think there are two different questions here. One is the roots of computational linguistics (and for that matter digital humanities), and the other is the development of the nomenclature (calling it "computational linguistics" or calling it "digital humanities"). And in fact, applications of digital processing technology to linguistic, to problems of language

and then also to problems in the humanities, is pretty much as old as the existence of computing machinery itself. So during the Second World War, the Enigma decryption project involved the construction of both special purpose and a movement in the direction of general-purpose computers, and although the goal of that technology was to decrypt transmissions, the transmissions involved were textual transmissions, and the calculations done had very much to do with the differential probability of different sequences of letters and words in German military transmissions. And so the problems involved, for example, the estimation of what we call N-gram models. The techniques commonly used there go back to ideas that were developed by Alan Turing in that project, so, you know, that was already in the early 1940s. So, you know, the term "computational linguistics" didn't really exist then, but the same algorithmic ideas were implicit in that very early work. And then, similarly, in the digital humanities, even back in the era of, you know, punched cards, sorting computing, one of the early applications that people worked on was the calculation of concordances of texts. I don't know exactly when that was first done, but it must have been done in the relatively early 1950s, I'm pretty sure.

[Federica]: So if I understand this correctly, it was computer scientists who first got interested in processing texts, rather than linguists got interested in the new technology and said, "How could we use this in our studies?"

[Mark]: I'm not sure that that... Again, I think there are there's the question of content versus the question of nomenclature and [organization?]. So one thing to keep in mind is that computer science as a discipline is actually relatively recent, and when... I mean, even as recently (not that it was that recently) as the time that I was an undergraduate, what we would now call "computer science" rather was carried out in departments of applied mathematics or in departments of electrical engineering. There were no departments of computer science; that came later. Similarly, linguists... So there was a lot of work in the, really starting in the 1930s and '40s, but becoming really more intense in the 1950s, along two lines that became and remain quite important in computational linguistics really begun by linguists.

So one was the idea of distributional analysis. So this is an idea that goes back to the structuralist linguists in America and Europe. It was particularly strong in the work of Zellig Harris. One of the slogans of that work in the area of syntax and semantics was that "You shall know a word by the company it keeps." That is, by looking at the distribution of word occurrences in large bodies of text, you should learn something. In fact, that would be the initial evidence from which you would infer the abstract analysis of morphology, syntax, and semantics. And as soon as computers became available for use, some linguists and computer scientists (actually, I suppose, electrical engineers and mathematicians in those days) began cooperating on trying to implement those ideas.

And then the second strand of work (again relatively early) was the work by Noam Chomsky and Pierre-Paul Schützenberger on what has come to be known as the Chomsky hierarchy or the Chomsky-Schützenberger hierarchy about a hierarchy of mathematical types of languages,

algorithmic processing automata, and types of rewriting rules in recursive function theory. So again, that was not... That was mathematics. It was presented as mathematics rather than as computation but has become a central part of the underpinnings of computational linguistics, although the people who devised it were a linguist and a mathematician, neither of whom actually worked on computers in those days. So I think there's a sense in which from the point of view of content, the marriage of linguistics and computer science and the marriage of computational analysis and humanistic processing of text and, for that matter, of audio, that existed as soon as it became possible to conceive of the relationship, and the nomenclature and the academic and industrial organization, the existence of scientific societies, engineering societies and so on — that developed on top, so to speak, rather than the other way around. It wasn't that there was computer science and there was linguistics, and then at a certain point they say, "Oh, let's form computational linguistics." Rather, people on both sides had their own problems that they were trying to solve in ways that involved computational methods. I hope that's not too complicated.

[Federica]: No, no, absolutely. I would just maybe like to ask a similar question from another perspective, and that is: Today, computational linguistics is an established field, so new generations of students can choose their programs at university and enroll in a program in computational linguistics — so how do you present the field to them today? How do you contextualize it in a historical perspective but yet give them a definition of what the field is today and also what kind of competences the students must acquire? How much tech savvy, how much of the traditional linguistics must they acquire? How is the profile of today's researcher or just expert in computational linguistics?

[Mark]: Well, there have been roughly three stages of development, and the first stage involved programming grammars and analyzers or parsers. So the idea was that we would figure out what the grammatical patterns of English, or French, or German, or Russian, or Chinese are, we would either write those down in a mathematical formal way, and then have ideally some kind of computer program that could interpret that formalism to either analyze or generate the relevant patterns or perhaps to do translation or something of that kind — or if that didn't work, we would write a parser as a specific kind of code that wouldn't interpret a grammar, but would look for certain patterns in strings of letters or would generate appropriate patterns. So that was stage one, which is human beings, computational linguists (linguists or computer scientists; it doesn't matter) writing grammars and parsers.

The next stage was the machine learning stage where, rather than writing the grammar and/or writing the code that would do the analysis, instead you would attempt to create a system that would learn the grammar and learn how to do the analysis using, for example, techniques of stochastic grammar, so it could be a stochastic finite state grammar or a stochastic context-free grammar, that is a grammatical formalism of a kind that computational linguists had developed, and it would go through lots of material and attempt by some iterative process

to figure out what the rules are and what probability should be associated with various options when things are ambiguous, which they always are. So that was stage two, stochastic machine learning.

And then stage three, which we're sort of still in the middle of now and the outcome is not so clear, is the applications of so-called deep learning, or deep neural nets, or pseudo-neural nets or whatever, which is a very general approach to pattern learning where typically very large amounts of material are put in, and rather than giving the system fairly detailed instructions about what it's supposed to learn (i.e., a grammar of a particular kind of form), instead you leave it open to the system (to some extent, at least) to form its own ideas (if we can call it that) not only about what the grammar is, so to speak, but even what a grammar is. That is, what kinds of patterns it's looking for. And pretty much the same kind of architecture might be used for recognizing pictures, or for recognizing speech, or for analyzing text, and that's kind of where we are now.

Now, I would say most students going into computational linguistics (whether relative to text, or to speech, or to both) actually probably learn things from all three stages, but the most common, practical, effective systems are now of the third kind, which, you know... There will undoubtedly be later stages, and one of the complaints that people have about this third stage, the deep learning stage, the most important complaint I would say, is: The systems that result are end-to-end black boxes. That is, you put in text or you put in audio and you get out the system's judgment about what the analysis is, but if you want to know why, there's nothing much to say other than, "This is what the system did." Whereas the earlier stages, the system processes the input and produces the output, but it also produces lots of humanly interpretable intermediate waypoints and structures.

A second problem with the current deep learning approach is that even though the same general kind of architecture can be used for image recognition, or for autonomous vehicle control, or for text parsing, or for speech recognition, or whatever, in fact, there are lots and lots and lots and lots of ways to modify that architecture that could be applied in any of those domains. Is it a recursive network? Is it a long short-term memory network? Is there an attentional mechanism? Is it a convolutional network? What's the batch size? What's the momentum? And so on. The meaning of those terms doesn't really matter. The point is that there are lots of piece parts, and knobs, and switches that have to be assembled and set in order to set up one of these projects, and it will work better or worse or maybe not at all depending on the choices you make, and there is no theory, mathematical or otherwise — no really effective theory — about how to make those choices.

One analogy that's been made is between these modern techniques and alchemy, because the alchemists were actually extremely effective practical chemists. You know, they knew how to extract acids, they knew how to create flavors and so on, but almost every process was kind of a thing in itself, so to become a good alchemist you had to engage in a very, very long apprenticeship, and even then when you approached a new problem, it wasn't really clear how to do it, because the theory behind all of this was either nonexistent or nonsensical.

[Federica]: That's a very nice analogy. I think I have one more kind of general question, so I could ask: What does a linguist do, and what does a computational linguist do that's different? But also, I add to that, is it more specifically the questions that are asked that are different or just the methods applied that are different?

[Mark]: Well, let's first make a differentiation between, on the one hand, science and maybe scholarship and on the other hand engineering and the creation of applications. And those, obviously, are different motives. So someone interested in science, for example, or in scholarship might be interested in, I don't know, how the verbal system of English developed from 1,000 years ago to today. No, that's not an engineering problem. An engineer might be interested in: How do we find names? How do we do what's sometimes called "entity tagging"? That is, how do we find in text names of people, places, organizations, dates and other sort of semantically well-defined — more or less well-defined — categories?

Someone who is interested in the history of English syntax, for example, people have been studying that before computers came into the picture by reading lots of texts and writing down lots of examples on file cards, and arranging the file cards, and counting things, and making either qualitative or quantitative judgments about how things changed. More recently, because we have almost the entire history of Middle English and Early Modern English available as digital text, we can use computational methods originally devised by engineers to go through and allow us to do literally in hours what previously would have taken decades.

And I could, you know, give you some specific examples — for example, work on the history of so-called "do support" in English, which was originally done by a linguist who did not attempt to use computers but just used old-fashioned, you might say almost philological, methods: reading texts and making notes — and more recently, a graduate student here at Penn who was able to use early English books online and some other sources of digital text and parsing technology in order to obtain several orders of magnitude more historical data not in a decade (as the previous work had involved) but rather literally over a period of days. And the value of the additional data is that he was able to look at the development over time of patterns with different verbs because he was able to get enough examples, decade by decade, of particular verbs in order to follow their trajectories separately. So the details there are complicated, and I won't go into them, but, I guess, again and again we find that there are matters of scholarship or science where we can use the techniques that engineers have developed and maybe extend them as well in order to address questions that are not, would not be of interest to the engineers qua engineers.

[Federica]: You have received your training in linguistics at MIT. You have your master's degree and PhD in linguistics there. During those years, Noam Chomsky was most active there. Was he one of your professors? Was he an important figure in your training?

[Mark]: He was certainly one of the people I took courses from, and I spoke with him quite a bit, and I think he was on my dissertation committee, for what that's worth.

[Federica]: You're an expert in speech and language technology. Speaking of technology, I would like to ask you if there is any special — well, piece of technology, a machine, something — that is only known or mostly known by the experts at the cutting edge of your research field that we non-experts might not know about that is involved in extracting data, processing data, some sensor maybe, and also if any of this is used in your research which I find very fascinating on the early diagnosis of Alzheimer's disease, so something that has an impact on health starting from the analysis of how a person talks. So I guess it's audio analysis.

Mark: Yes. Well, audio and/or text. So, again, there are both scientific questions and technological questions in the area of clinical applications of language — speech and language — analysis. So let's take another case which I think is even clearer — namely autism, or, as it's sometimes called, the autism spectrum. And this is an area whose definition is constantly changing. The most recent DSM made some very serious changes in what counts as autism and what doesn't, and there are a bunch of other related characteristics, related categories, like social anxiety disorder, for example, or ADHD. And people, for quite a while, rather than talking about autism as a well-defined box you can put people in, have started talking about a spectrum where people can be arranged along a line in some sense. I think anyone who looks at this problem quickly realizes that it's not just one dimension, that it's not a spectrum but a space, and furthermore, it's a space that we all live in. It's just that some corners of the space, because they interfere with people's ability to carry out ordinary life, have been sort of medicalized.

So one scientific question is what the dimensions of the space really are, and some have to do with language and communication, and some have to do with the ability to understand other people's goals, and intentions, and beliefs, and knowledge, and others have to do with interests and preferences and so on. There are many aspects. And it's likely, I think, that if we had a large amount of the right kind of evidence where computational linguistic and speech analysis could help us with that, that we could really start figuring out what the true latent dimensions are as well as, importantly but I think scientifically less interestingly, developing diagnostic techniques for placing people in this space.

And I've done some work along with colleagues at the Center for Autism Research at Children's Hospital here in Philadelphia aimed toward that kind of goal. Autism, obviously, applies to people throughout the life cycle. There, the issue that's especially important is, how early is it possible to make a diagnosis of one kind of problem or another in young children? Because there are interventions, mostly behavioral interventions, that really can help, and you want to start them as early as possible, but you don't want to start them unnecessarily. So one of the questions is sort of what are the dimensions of variation that are relevant, and what are manifestations that might appear in young children, maybe even before the age of one year?

At the other end of the life cycle, of course, there are the neurodegenerative disorders like

Alzheimer's disease, which pretty much all of us are very likely to suffer from if we live long enough, and there again, there's a hope, at least, that if we could detect the onset earlier, then there might be therapies that would help. At present, there don't seem to be any, but early diagnosis is, again, a sort of interesting possibility.

Even more important, I think, is tracking the time course. So suppose that we have a drug that we believe might, in some or all cases, slow the onset of Alzheimer's or even reverse the syndrome. How do we test that drug? Well, we pick perhaps a few hundred subjects who are at risk of the disease and we divide them into the clinical group and the placebo group, and we give them the placebo or the medicine, and now we do something over a period of six months, or a year, or two years. But what is it that we do? How do we determine whether they're getting worse, or getting worse at some rate, or getting better, or staying the same? But we need some metric, and we would like a metric that is relatively non-invasive, that is relatively inexpensive, that is relatively low-stress for everyone concerned, and it seems quite likely that a speech-and-language-based metric will actually fit those needs very well.

[Federica]: In the early '90s, you founded and you're still director of the LDC, the Linguistic Data Consortium. What is this about, what is it for, and how is it doing today?

[Mark]: Well, taking those questions in reverse order, it's doing fine. What is it for? The original issue was that, starting in the mid-to-late 1980s, some people in the U.S. government, in U.S. government funding agencies, decided that it was important to try to fund engineering research in speech and language areas. The two most important ones, though there have been many others, were automatic speech recognition and machine translation. Others include things like language recognition, speaker recognition, and a bunch of other things as well, information extraction from text or from speech, document retrieval and so on. And that kind of work, as of 1985 let's say, had a very, very bad reputation because people had begun working on it with something in between naive optimism and maybe a little bit of dishonesty since the late 1940s, early 1950s, making big promises about what they were going to achieve. There was a book published in the early 1950s called Giant Brains which projected, for example, you know, things like the early ENIAC computer and so forth, that within a decade or two computers would be housed in buildings the size of the Empire State Building and would [unclear] the entire electrical output of Niagara Falls. But also they felt that within a decade or so, a voice typewriter would exist. And of course, the computer size thing turned out to be totally and completely in the wrong direction, and the voice typewriter, one might say, has arrived, but it took, you know, 70 years, 60 or 70 years, not 10. And there were many, there had been many failures, people who promised that they could achieve machine translation or they could achieve speech to text and who really didn't deliver much. And so there was a period of a decade or so that people sometimes refer to as the AI desert when at least U.S. government funding in those areas was almost completely withdrawn, and in fact most commercial, most industrial labs also backed away or entered very gingerly into those ideas. I experienced that because I got my

PhD in 1975 and went to work for AT&T Bell Laboratories, and the idea of working on speech recognition was something that people were very leery about, and they would make very small promises and try simple things like, say, isolated digits or maybe digit string recognition, not a general voice typewriter. So anyway, as of the late 1980s, the government decided that in order to get into this area, they would need to carefully guard themselves against, the founder of my center at Bell Labs, John Pierce, called "glamour and deceit, otherwise known as BS." And in order to do that, they wanted to have a series of very well-defined tasks with well-defined examples eventually becoming training material, very well-defined test material, and automatic, well-defined evaluation programs that could be carried out, implemented by the National Bureau of Standards, National Institute of Standards and Technologies now. So that was the way things were set up. The problem that immediately arose was how to organize the creation, curation, and distribution of these large bodies of digital audio, and text, and video, and images to some extent as well. Remember, this was in the late 1980s. Google didn't exist. The internet didn't really exist. The main way to send things around was on... I don't know if you've ever seen a 9-track tape, but these are large 12-inch platters of magnetic tape that maybe contain a megabyte or so of information each. So it was a different world. But there also were problems of intellectual property rights, of human subjects' permissions, and things of that kind. And anyway, they started trying to do this through the National Archives and other government agencies, and in those days those, at least those agencies were not really set up to do this kind of thing and not all that interested in doing it. So anyway, Charles Wayne, who was then the Director of Speech and Language Research at DARPA, decided that it would be a good idea to have an organization housed in an academic context that could take care of those things and maybe also encourage a kind of marketplace, you might almost say, or, you know, open publication of materials not just from these government programs, but from other kinds of research entities. And so he got some funding from Congress to set this up, and I joined with some other people in 1987 or 8 in drawing up a white paper about how this might be done, never thinking that I would be involved in actually doing it.

In 1990, I left AT&T and came to Penn as a faculty member, and I was very happy to put research administration behind me and return to research and teaching, but Charles asked me to apply on behalf of Penn to be the home of this Linguistic Data Consortium, and I was only able, effectively, to say "no" to him for about three months and eventually gave in and did it. And this was, what started out as the top right-hand drawer of my desk has now turned into the floor of an office building with 50-odd employees and lots of computers.

[Federica]: Basically, this consortium provides linguists all over the world with materials and I would imagine today large datasets for their experiments and analysis, like a coherent, common, shared, open, public corpus of data?

[Mark]: Yeah. The datasets are of different sizes. Some we create. About half or a little more we publish on behalf of other people. Sometimes we publish things on behalf of IBM

and Google and other companies, but we've also published on behalf of academic institutions all around the world — in the United States, in Europe, in Asia. And what we do, whether for our own material or for other people's material, is, we create and maintain documentation and catalog entries, we do quality control. In some cases, we have to do more. Sometimes, what people send us is, you know, a cardboard box full of analog tapes and some typescripts or something of that kind. We arrange for normalization of formatting, and then we do curation. That is, we produce later editions if there are corrections to be made or additions. We handle intellectual property rights, negotiations, and make sure that privacy and confidentiality and human subjects constraints are obeyed and so on. Or. . . I say "we." I mean, it's, of course, the people who work there who do it.

[Federica]: In the early days of computational linguistics, although it was not called so, text processing was the main thing. When did sound processing become being a thing, like speech, etc.?

[Mark]: Well, I don't know for sure, but there was already digital audio in process in the late 1930s, early 1940s. Actually, there's a literary connection. So if you read Solzhenitsyn's novel The First Circle, it's about a laboratory near Moscow which is a fictionalized version of a laboratory he actually participated in which is staffed by political prisoners and whose goal is to do what I guess we would call computational speech and language research, mostly speech. And in particular, the novel deals with two technologies. One is speaker recognition from audio recordings — obviously, in that context, intended for, you know, purposes of political repression — and the other being encrypted telephony, encrypted voice transmission. And the encrypted voice transmission is something that was developed jointly in England by people that included Alan Turing and in the United States at Bell Labs by people including Claude Shannon, and they collaborated on this. And it was actually implemented in such a way that Churchill and Roosevelt could use an encrypted radio transatlantic telephone conversation, basically. Now, this involved, you know, a room full of complicated and power-hungry apparatus in London and one in Washington, but it did involve digital voice, as I understand it. And, anyway, in the sharashka that Solzhenitsyn writes about, they were attempting to build something like that for Stalin because Stalin was jealous of the fact that Roosevelt and Churchill had it and he didn't. And one of the things that's clearly there is how to develop technologies for producing speech to bits, doing things with the bits to provide encryption, decrypting on the other end, and then reconstituting the speech. Now, the novel is not primarily about technology, but the technology is there in the background and I think is actually quite accurately portrayed.

[Federica]: You've been in this field for many years, and I'd like to ask you something about technological evolution. Have you witnessed something during your career that came from the technological front that was remarkable, that had an impact on the way you do research in this field, an advancement that's significant enough to share?

[Mark]: Well, frankly, the most remarkable changes are the result of the same developments and forces that are changing everything else in modern life — namely, the development of ubiquitous, inexpensive, high-bandwidth digital networking, the exponential improvement in the cost performance of various kinds of computational devices, including computers and various kinds of wearable devices, as well as the cloud (as we call it now), and also from the point of view of speech and language, especially the incredible changes in the cost performance of mass storage. So it's been a while now that arbitrary amounts of text are in effect trivial or free to store. That is, you could take what were 20 or 30 years ago unimaginable amounts of digital text and put them in your shirt pocket. And speech is rapidly moving in that direction. That is, again, for a relatively small amount of money, you can buy a mass storage device on which you can put enormous amounts of audio, along with transcripts if they're available, and so on, and, you know, you can download tens of thousands of hours of audiobooks from the web, and you can go to YouTube or other places and see millions — probably many more than, many millions — tens of millions — of hours of speeches and songs and readings and so on. And, you know, for anyone interested in analysis of speech and language using digital means, it's like walking into an amazing magical garden.

[Federica]: That's a very nice image, again. So it's nothing in particular, but the development of technology itself that is what's fascinating.

[Mark]: And I think the most important, the most impressive, the most valuable, the most interesting and insightful developments are actually still in the future.

[Federica]: Speaking of the future, I would like to ask you to share a vision for computational linguistics in the future — not necessarily a realistic one, just an optimistic vision that you have of where this research field could go. In the best-case scenario, how do you see computational linguistics in 5 to 10 to 20 years?

[Mark]: Well, there are many kinds of speech analysis where we start either just with the audio or perhaps with the audio and a transcript, but it's much, much easier, obviously, to create audio than to create audio with transcripts, just because transcription is a somewhat labor-intensive process and fairly expensive. So at some point in the next, I would say, 20 years or so, speech recognition will get good enough that across a wide variety of kinds of input for a wide variety of languages, we'll have speech-to-text which is good enough to be able to do away with most cases of human transcription. There have been some claims that we're there now. Several companies have claimed human parity (as the expression goes) in speech-to-text transcription, and they have actually achieved something like that in particular domains, but it's not the case that across arbitrary kinds of content, arbitrary kinds of recording conditions, arbitrary kinds of modes of interaction and so on, backgrounds and whatever, that automatic

methods can reliably work. They work very well sometimes and they fail completely in other cases, but 20 years from now, that won't be true.

The next factor, the next thing there is that once we have the transcript and the audio, we can do what's called forced alignment, and we can figure out quite accurately which words occur where, but we do that despite the fact that people are not dictionaries. That is, even someone who speaks the standard variety of whatever language we're looking at doesn't pronounce words in spontaneous speech the way the dictionary says they should. Sometimes they do, but more often there are various forms of lenition, or reduction, or modification, and we have specialized ways of looking for those effects in particular cases in particular languages for particular kinds of material, but we don't have sort of what one might call an automated phonetician.

[Federica]: I ask for your forgiveness, but I'm not quite following. What's so extraordinary in speech-to-text recognition? It seems to me that I missed the application. What happens next?

[Mark]: Oh, okay. Well, let's take an example. We're working in this Alzheimer's and other analysis area, or at least that's the goal, on a body of material from the Framingham Heart Study, which is something that the national center for, National Heart, Lung, and Blood Institute in the United States began in around 1950. They recruited pretty much as much as they could of the entire adult population of a small town west of Boston — Framingham, Massachusetts — and they began keeping track of medical history issues and lifestyle factors and other things for all of those people over time in order to try to figure out what the factors were that influenced coronary disease. Around 20 years ago, they broadened their scope to include neurodegenerative disorders including stroke but also neurodegenerative diseases, and so they began giving a battery of neuropsychological tests which lasts about one to two hours to each of the members of their original cohort and the cohorts that they've recruited since then. And since 2005, those neuropsychological test batteries have been, the audio has been digitally recorded, and the results of the testing are graded by the test giver, by the interviewer, with pencil and paper at the time of the test, but there's a lot more information in the audio that isn't captured in that grading which they've never done anything with because these things are not transcribed. So as of a couple of months ago, we got about 10,000 hours of these interviews collected over 15 years or so, and we've begun a process of transcribing a selected set of them. Now, at best, we can hope that the human labor to transcribe an hour of this audio is about 10 hours, so for 10,000 hours of audio, we're talking about 100,000 hours of human labor. It would be very nice not to have to do that, and doing that is the first step toward what we really want to do, which is looking at things like latency to respond, speaking rate, where, rate of disfluency and where disfluencies occur relative to types of words etc., etc., etc. Once we have an accurate transcript and the audio and we can align them, then those things can be worked out, although to work them out in detail, we would want more than just what words occur where. We would want to know how those words were pronounced, and again, that's

something that that would be another layer of automated analysis. So at present we can do all of those things with human labor; it's just expensive and time consuming. And if you wanted to put this into effect as a kind of automated diagnosis or classification or screening method, you would probably ideally would want to be able to automate it rather than having humans in the loop.

[Federica]: Is one of the main obstacles in text-to-speech recognition the gap between the language as it should be according to the rules and how people actually speak with broken sentences, slang and accent?

[Mark]: Really not. That in and of itself is not necessarily a problem. So among the bodies of material on which various groups have achieved something like human parity in transcription are spontaneous telephone conversations on assigned topics between strangers, but still. You know, these are not professional speakers. They're ordinary people. They have all kinds of different accents. They don't all use standard English. There's certainly plenty of disfluency, plenty of slang. All of that is learned as part of the language model for that material, which is one reason that techniques are able to work fairly well. No, I think there are two key areas of problem... What you refer to is a problem, but the biggest problems are, effective speech-to-text very much depends on an accurate idea of what people are likely to say or not say. And so if someone is, you know, telling a story about one topic, versus arguing about a political issue, versus, you know, reading a translation of the Iliad, versus describing what they see outside their window or whatever, the language model, the question of what they're likely to say or not say varies a great deal quite independent of whether they're using slang or standard language. So that's one thing that is just variation in the a priori language model. And then the second thing is recording conditions: reverberant speech, background noise, music in the background, multiple people talking at once. Those things, at present, are very, very hard for these systems to deal with effectively.

[Federica]: We've mentioned that you're a speech and language technology expert. What is the difference between speech and language? What is the technical definition of one as opposed to the other in your research field?

[Mark]: Well, of course, versions of this distinction have a long history. There's Saussure and langue and parole and so on, but crucially: "speech" is obviously talking (and that could be spontaneous talk, and it could be baby babbling, it could be somebody reading a book, and it could be a political debate, could be lots of things); "language" is a term generally used for the more abstract system, you know, which includes things like what are the words, and what do they mean, and how are they inflected, and how are they arranged in phrases? And the language could be instantiated in speech, or in text for languages that have a literary culture. So "speech and language" is just a kind of a cover term meaning all that stuff,

and you're absolutely right that speech involves language. It's just that sometimes when people talk about studying language, they're really thinking about studying text more than speech.

[Federica]: If I understand correctly, it's common practice today in computational linguistics to consider both text and audio or the transcription of audio. So can you give an example of a type of study that you can do with these two sources combined together?

[Mark]: Well, pretty much almost any kind of phonetics. There are kinds of phonetic analysis where you're only concerned with something like, you know, the distribution of fundamental frequency values or something, but mostly you might be interested in, you know, what's the range of ways in which someone pronounces a certain vowel? Sociolinguists very often study changes over time or over space in vowel quality, especially in a language like English where those changes are ubiquitous. In Spanish, people often study the lenition or deletion of syllable final /s/ and the lenition of intervocalic /d/, so someone who instead of pescado says pecao, as people might in some varieties of Spanish. And so on. So pretty much anything that involves, for example, looking at changes over ethnic group, space and time, of the pronunciation of particular vowels, consonants, either in general or in particular contexts, and how that relates to overall change in the language, that's something for which you need the transcript as well as the audio. I mean, one of the things that has happened to help out sociolinguists is, they used to have to make tape recordings and go through and listen to them and then, you know, do analyses of few vowels or consonants that they found here and there either by their subjective judgments or by making measurements, and now you can put in hundreds of hours of speech, put in the transcripts, do the forced alignment, and automatically pull out the measurement of the formant frequencies or duration and spectral centroid for the s's or whatever for everything in that data.

[Federica]: 2018 is the European Year for Cultural Heritage, and heritage is also about identity, and language has a lot to do with it, so I would like to ask if in your community there is an awareness of the importance of preserving these recordings that you have of various types of speech also because, in the future, they will become ever more valuable, so it's like a way of preserving accents but also life stories.

[Mark]: Well, so let me say two things. One is that language preservation and documentation is obviously an important and growing field, and one aspect of that involves small languages that are dying out or at risk of dying out because fewer and fewer people speak them, and because speakers may not feel that they have a path to the modern world, and so younger people, although they may understand, no longer speak and then their children don't even understand. But something that's less well recognized (although there certainly are people in the U.S., in Europe, and elsewhere who do recognize it) is that there are endangered languages and varieties all over the world, including in Europe and in relatively wealthy countries, like growing

countries like China. So, for example, in the Netherlands there are local varieties of Dutch, some of which I understand, for example, the variety spoken in Rotterdam is fairly lively, but there are many other varieties that were just spoken in a small town or a range of villages where young people don't speak it anymore and the old people who speak it are dying out, so one could almost argue that the density of endangered varieties, at least, if not languages, is actually greater in Europe than almost anywhere else in the world, because there are all these local varieties of Dutch, and German, and Italian, and Spanish, and Czech, and Russian and so on, even English and Scandinavian languages, which will almost certainly not exist in their current form in 100 years and may not even exist, in some cases, in 10 or 20 years. So I think [there's] really quite a bit of urgency as well as a possibility to do that kind of documentation.

[Federica]: It strikes me that you use Europe as an example of a place with a great variety of languages, because of course there are many languages in Europe and the dialects too, but when I hear how many languages there are in the entire world, it seems to me that Europe only counts for a small fraction of that, and that there must be other areas in the world that, for many reasons, have even a greater variety even locally, like from this village to the next village 10 kilometers from there.

[Mark]: Oh, absolutely. I mean, before nation states come along to change this, the state of nature so to speak (especially in settled agricultural civilizations) is generally for a kind of geographical dialect continuum. I mean, it has been said that there was a time when you could walk from the English Channel to the Straits of Gibraltar, you know, through France and Spain or from the English Channel to Sicily through France and Italy, and never pass a point where the people in one village couldn't speak with the people in the next village. That is, it wasn't so much that there was French and Italian and Spanish as well-defined, strikingly distinct varieties; there was just this continuum of Romance varieties. And, of course, as nation states decided, especially after the French Revolution, that it was important to have a standard national language that everyone learned — and also, more controversially, important to suppress all the other varieties — that has begun to dissipate. My impression these days is that because of the pressure of various kinds of modern opportunities and modern cultural forces, it's no longer primarily state actors that are suppressing local languages. It's rather the opportunities that are involved, afforded to those who become fluent and primary users of the standard language.

[Federica]: How many languages are there in the world, actually? I know it's a bit of a stupid question, again, but we take this for granted. We just go online and then find a number, but who does actually count? Who goes around and counts the languages, and who keeps track of the ones that are dying out? Who labels those which are endangered? Is there an official register with a complete list of languages?

[Mark]: Well, the people who first began doing that are SIL (Summer Institute of Linguistics), who created some time ago the Ethnologue publications and online resource. They're a Protestant Bible translation group, basically, and they began collecting this so that they could figure out which languages needed Bible translations, but I think they have a genuine interest in it as well. More recently, there's an International Standards Organization, ISO standard, that accepted SIL's three-letter identifiers for languages and varieties. And so I guess if anybody is keeping track of this, it would be SIL and the International Standards Organization. But obviously, the question of what's a distinct language and what's a variety is not an easy question to answer. In fact, it's not really a coherent... It's not a question that has a coherent answer. As the famous saying goes, "A language is a dialect with an army and a navy," from one point of view. It's a political question, not a... I mean, that's no longer entirely true, because there are plenty of things that are recognized as languages that don't have independent nation states, but it remains the case, I think, that it's a political question as much as a scientific or linguistic question.

[Federica]: Yeah, something that was a bit implicit in my previous question is that a language needs also to be recognized, in fact, so dialects are not always considered languages, and this matters in the total count of how many languages we have in the world. I know that at least in Italy, there are two... Huh, well, what do you call them? You could call them "dialect," but they are in fact recognized as languages, so they are languages. They get pretty mad if you call them dialects, but there must be, you know, a way to transition from status of not being recognized to being recognized, so how does that happen? Are there linguistic groups who at some point apply, to whom, to be recognized as official languages?

[Mark]: There are certainly political movements. For example, after the breakup of former Yugoslavia... So there was a time when Serbo-Croatian was the name of a language, and it was recognized that there was a continuum of dialects across Serbia and Croatia, and it was obviously known that the Croatians use Latin characters, the Serbs use Cyrillic characters to write their language, but that's just how you write it; it's not what the language is. But after the breakup of the former Yugoslavia, there has been a lot of pressure to say, "No, no. There's Serbian and there's Croatian," and to purify Serbian and to purify Croatian — in the case of Serbian, by removing German borrowings; in the case of Croatian, by removing Russian borrowings. I don't know how successful, from the point of view of changing the way people talk, those efforts have been, but they certainly result in the fact that people no longer talk about Serbo-Croatian as a language.

[Federica]: I would like to thank you very much for your time. I'm very glad we spent this hour together. I have personally learned a lot, although we have not covered all the aspects and all the possible implications of the research that you do, but I really appreciated this, so thank you for being on Technoculture.

[Mark]: Thank you very much. It's been a pleasure, and I look forward to hearing from you.

[Federica]: Thank you for listening to Technoculture. Check out more episodes at Technoculture-podcast.com, or visit our Facebook page at technoculturepodcast and our Twitter account, hashtag technoculturepodcast.