Jon Krosnick: I should just, for the recording’s sake, Mark Liberman thinks of himself as very outside this group because he is a linguist. But we in Survey Research are very excited about the fact that he and some colleagues are intrigued enough and interested enough in the open-ended data that we’ve been collecting, and it might serve some of their purposes such that they are going to be analyzing some of these ANES data with machine-coding approaches, and so, over the coming months, we’ll hear what they come up with.

But we’re really glad that Mark was willing to come today, into foreign waters, and talk with us about how computer-based natural language processing and HLT work, and what their – what the potential might be for survey applications. So, thank you, Mark.

Mark Liberman: Hello, okay, great. I’d like to thank Jon and the other organizers of this workshop for giving me this opportunity to have a cross-cultural immersion experience, which I’ve already benefited from considerably; I’m beginning to feel some osmotic changes. So, I’ll try to return the favor to some extent, to the extent that I can do so as a single participant.

So, HLT is a term – is an abbreviation of ours that has come to be used especially in the DARPA Speech and Language program, starting in the mid-1980s. It’s a little bit more general than just smart things you can do with text. It includes also smart things you can do with speech in and out. It’s a bit broader than natural language processing, computational linguistics, speech technology, and a number of other rubrics that such work also often goes under.

What it means – connects to lots of different individual technologies: document retrieval, which is what googling used to be called; document classification; document understanding; information extraction from text, which is less – a less grand goal than understanding a document, just finding certain entities in relation, say, in it; summarization; question answering; what’s come to be called sentiment analysis, or sometimes opinion mining, and so on. Machine translation, I guess I shouldn’t ignore because it’s become increasingly important.

And then for speech, there is speech recognition, where speech goes to text; there is speech synthesis, where text goes to speech; there’s the recognition or classification of speakers; what’s called diarisation in the trade, which basically means figuring out who spoke when, if anybody did; spoken document retrieval; information extraction from speech; question answering; human-
computer interaction via speech, ala Siri, of example; speech to speech translation, and so on again.

Now, how I think people 50 years ago, and many people still today feel that human language technology ought to work is that there’s some kind of human interaction, whether spoken or written, and there’s some kind of somewhat mysterious artificial intelligence that has the general ability to understand what people say, what they right, what they mean, why they meant it, and so on, and is able to put that into the form of some data that computers can process, and the arrows go both ways because now the data can come back and form answers to the people involved in the interaction and so on.

In fact, the way that it really does work, or has come to work, you could assign those pictures to these boxes. But if we look at what’s actually in the boxes, there’s some sort of input: speech, or text, or video. Then there’s some extraction of features, and this is usually not terribly mysterious. It could be time functions of spectral features in the case of speech. It could be words and word sequences in the case of text.

There’s also some human annotation: transcription, classification of documents, finding regions of text that correspond to certain classes of entities: people, places, times, dollar amounts, governments, genes, proteins, disease states, whatever.

And then there’s a statistical model of some kind. It could be a very simple statistical model. It could be a very complex and arcane and difficult to understand statistical model, but it’s basically just a model of this array of inputs, which includes the automatically extracted features, and then the human assigned features.

And then in the testing or operational phase, in effect, the same thing is happening, but now the transcription, or classification, or other kinds of things that were previously done by humans are actually generated from the statistical model and the automatically extracted features.

In a sense, you can see this just as a form of regression, that is there’s some independent variables. There’s a statistical model. There’s some dependent variables, which in the training phase you learn parameters for the model that relate the inputs and the outputs, and when you run it, then you just want to get the outputs automatically.
Now, this sort of – applications of these ideas in survey methodology have been around for quite a long time. I don’t actually know how far back it goes, because I don’t really know the survey field, but it was easy for me to find a 1998 book, which is not – this is not very sharp, because it’s based on a screen grab that I got from Google Books, but this part of the then diagram includes processing a response to open-ended questions.

And in fact, one of the reviews of this book, in a computational linguistics journal, complained that, “There’s all this stuff in there about surveys, and who cares about surveys,” basically is what they said. You know?

[Laughter]

Mark Liberman: “Why don’t they do some interesting stuff?”

Okay, so one question is why has there been, in the now 15 years or so since 1998 – and in fact, I think this stuff was going on 10 or 15 years before that – why aren’t all of you using this stuff every day?

So, maybe the human language technology applications didn’t really work well enough. That’s certainly one possibility. And so, we might want to ask, to the extent that that’s true, has there been enough progress to change this?

And, of course, they could fail to work for several reasons, either because the basic algorithms are insufficient, or because there were implementation problems, or whatever other reasons. Maybe the two cultures have been too different, and I think the two cultures are quite different.

People don’t necessarily talk to one another, even when they’re at the same institution. I learned, for example, that Jon, when I first came into contact with him, didn’t know my old friend, Dan Jurafsky, who’s a computational linguist at his – at Stanford, at his institution. He’s been there for a while and is one of the world leaders at finding applications in this area, and who’s joining us in looking through the ANES data.

And so, one thing we might ask is, is this changing, or will it change? Maybe should it change? Are there reasons for increased interaction, and what are the reasons, and what kinds of interaction make sense?
So, there is a key question here, which is how well does all this stuff work now, and it really still depends. It depends – there’s nothing mysterious about this, frankly. The details can be arcane, but it’s like any other attempt at statistical prediction; it depends on the appropriateness of the model, the amount and quality of the training data, and the inherent difficulty of the particular task.

As a result, out-of-the-box solutions to one problem or another will sometimes work to solve survey research problems, but not always. And the process of figuring out whether or not they do may be somewhat protracted and difficult, as I think the people who spoke before lunch made clear.

I wanted, actually based on Arthur Lupia’s presentation, to say a little bit about our – the experience over 25 or 30 years of research in human language technology on human coding, which is that natural annotation is extremely inconsistent. If you give annotators a few examples, or a simple definition, and turn them loose, what you get is very poor agreement. Now, this is from a presentation I gave that was talking about identification of entities, which are things like people, places, organizations, or, in a biomedical domain, genes, gene products, proteins, disease states, organisms, and so on.

So, even – you know, take a bunch of Ph.D. geneticists, give them scientific papers about – that are about areas in their specialization, ask them if they can determine when genes are mentioned in those papers, and they’ll look at you like you’re an idiot. Take two of them, put them in separate rooms, and ask them to do that task, and then look at how well they agree on the output. If they agree 50 percent of the time, you’re very lucky.

It’s worse for what we call normalized entities, that is where you don’t just say, “All right, this is the name of a gene,” but you want to know which gene is it the name of. This is – not just, “This is the name of a politician,” but which politician is it the name of?

Worse yet, for relations among entities. This is because human generalization from examples to principles is variable, and human application of principles is equally variable. And the natural language context raises a bunch of additional hard questions. The result is that the “gold standard,” as we call it, is not naturally very golden. The resulting learning metrics are noisy, and F-score, which is the harmonic mean of precision and recall of .3 or .5, it’s just not a very attractive goal. If you tell people that you can agree
with their intuitions 30 percent of the time, they’re not very impressed, even if their intuitions and their neighbor’s intuitions only agree 30 percent of the time.

So, the traditional solution is an iterative refinement of guidelines of exactly the kind that we heard about before lunch, try some annotation, compare and contrast, adjudicate and generalize, go back to step one and repeat, at least until inter-annotator agreement is adequate.

What we usually do is about ten percent blind dual annotation throughout the task. That is about one instance – one thing to be annotated in ten is actually being annotated by someone else, and you don’t know which ones those are.

The process of convergence is slow. We heard that it took them four years. We can sometimes get it down to months, but that’s usually because DARPA insists that we deliver the data, not so much because we’re really convinced that we solved the problem.

Now, the result can be quite high inter-annotator agreement. For things like entities, we typically get over 90 percent. But this is based on a complex accretion of what is really quite like the common law. It’s like you have a simple statute that says what you’re allowed to do and what you’re not allowed to do, what’s legal and what’s not legal. Then when you try to apply that to cases, it gets complicated. And if you want to do it in a consistent way, the only way that seems to work is to accumulate a long list of semi-generalized particular examples.

Our guidelines for typical annotation tasks can sometimes be hundreds of pages. They’re typically at least 50. They’re slow to develop and hard to learn, but this is more consistent than natural annotation. It’s the only way that we know to produce high-quality inputs to the machine learning process that could be used to automate, as we learned before lunch.

Okay, now getting back to the potential for interaction, the simplest reason to want to do it is that there’s a not inconsiderable overlap between what surveys do and what human language technology or human language technology researchers do.

For example, open-ended response classification is effectively equivalent to spoken document – to document classification, whether written or spoken. And this was the overlap that was featured in that diagram from the 1998 book.
I think that there’s a larger and maybe better set of reasons – well, saving money and getting things done faster and to a higher standard is a very important goal, and so I don’t mean in any way to minimize it. As an academic researcher, I guess I’m always interested in the things that we haven’t done yet.

And there’s a large area of what surveys could do, and some, I think, do already, that also overlaps with interests and applications in human language technology. Probably the most obvious one, and the most widely explored these days is sentiment analysis or opinion mining and perhaps its generalization to things like trying to figure out how strongly respondents believe what – the answers that they give you, and therefore, how likely they are to change their minds, that sort of thing.

I thought I would give you a fun – an example of a fun, easy case that we did in a class last semester. So, I got 84,000 reviews. One review is from Wine Enthusiast Magazine online, which has about 4 million lexical tokens corresponding to around 30,000 distinct words, of which about 5,500 occurred at least 20 times. Each of those reviews comes with a rating, which is a number between 80 and 100, which is supposed to be how good the person who created the wine-tasting notes thought the wine was.

So, we did a regression of the ratings on the words, limiting it to the 5,500 words that occurred at least 20 times, and the multiple R number for the regression was .86.

So, about 75 percent of the variance in the ratings is accounted for by this simple bag of words model. And you can sort of see why that is, if you look at the 20 words with the biggest regression coefficients, “incredibly,” “gorgeous,” “superb,” “brilliant,” “beautiful,” “wonderful,” “massive,” “wonderfully,” beautifully,” “opulent,” “delicious,” “impressive,” “powerful,” “excellent,” “luscious,” “huge,” “power,” “intense,” “long,” and “richness.”

And similarly, if you look at the 20 words with the smallest coefficients, “heavy,” “lean,” “sour,” “modest,” “everyday,” “rough,” “rustic,” “simple,” “short,” “dimensional,” which presumably comes from one dimensional, I would guess, “thin,” “vegetal,” “watery,” “lack,” “lacking,” “dull,” “lacks,” “harsh,” “sugary,” and the one that’s interesting to me, the single word with the most negative regressions coefficient was “acceptable.”

[Laughter]
Mark Liberman: That’s like the worst possible thing for a wine to be apparently.

Audience: As long as it’s bad, it won’t get into –

Mark Liberman: Well, yes, probably that’s true.

I’m going to skip this for – in terms of time. And I think there’s an even larger and even better set of reasons, which is that there are things that researchers in human language technology would like to do that are somewhat different from, but also overlap with possible future survey research, and most importantly would benefit from access to the kind of datasets that survey researchers have.

So, what will the future bring? So, as I said before, the out of the box solutions, some of them, I think, may work now. As time goes on, more and more of them will work, but in my opinion, the biggest potential benefit is in collaborations where both sides are breaking new ground. And I’d like to say a little bit about how that could work.

And the best way that I can explain this to you, I think, is by telling a story. And so this is my attempt at intercultural communication from the other side. I want to give you some insight into an interesting and perhaps, from your point of view, slightly peculiar feature of researchers in this area, and the culture of that field.

So, the story begins in the 1960s, with some interventions by John Pierce, who was an executive at Bell Labs, who invented the word “transistor.” I think he supervised the group that invented the transistor, and he also supervised development of the first communication satellite. This is what he looked like. It was his organization in Bell Labs that I got a job in when I got out of graduate school.

So, in 1966, he chaired a committee, assigned by the National Academy of Sciences, to report on research in machine translation in automated translation of text. And in 1969, he wrote a letter to the journal of the acoustical society of America.

[Break in audio]

Mark Liberman: The ALPAC report, which was the machine translation report, was diplomatic.
“In 1966, machine translation was not very good, and ALPAC said that the committee cannot judge what the total expenditure for research and development should be; however, it should be spent hardheaded toward important, realistic, and relatively short-range goals. And in fact, U.S. government funding for machine translation research went essentially to 0 for more than 20 years.

“The committee felt that science should precede engineering in such cases. We see that the computer has opened up to linguists a host of challenges, partial insights, and potentialities. We believe that 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.”

So, this was 1966, and it’s – we don’t yet have the quantum field theory of linguistics, but I guess they don’t have a quantum field theory in physics either.

Pierce’s views about automatic speech recognition were similar, but his letter was his own personal composition. It wasn’t a committee report, and so it was substantially blunter.

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

“Most recognizers,” by which he meant 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, or acceptance of untested rules, the untrustworthy engineer approach.

“The typical recognizer builds their 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.”

And then he 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,” which they did after – anyway.
Mark Liberman:

“One doesn’t attract thoughtlessly-given dollars by means of schemes for cutting the cost of soap by ten percent. To sell suckers, one uses deceit and offers glamour.” Now, this is pretty – this is – he’s not mincing any words here.

“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.”

So, the first idea at the government level was, “Okay, there’s this new notion that’s come along, artificial intelligence, people that – very smart people at MIT, and Stanford, and CMU, and various other places have come up with all of these brilliant ideas about applying logic and proof theory and stuff like that to solving problems, so, let’s try it.”

So, there was a DARPA speech understanding research project, ’72 to ’75, which tried out those ideas, and it was canceled early on the grounds that it didn’t seem to be getting anywhere. Basically, the Piercians in the government funding agencies won.

And the second idea was just to give up. So, between ’75 and ’86, there was no U.S. research funding in these areas at all. There was a fair amount of work in Europe; there was a certain amount of work going on in Japan; some work going on in U.S. companies, though relatively little, but nothing really funded by the U.S. government – hardly anything.

And he was not the only person with this jaundiced view of this area. By the mid-1980s, most – many, anyway, and maybe even most informed American research managers were skeptical about the prospects. But there were also many people who thought that some kind of human technology – human language technology was needed, and in principle, ought to be feasible, at least to some level of performance.

So, in 1985, the question was posed, “Should the Defense Advanced Research Projects Agency restart human language technology?”

And Charles Wayne, the DARPA program manager, had an idea, which was to design a speech recognition program that would
protect against glamour and deceit, because there’s a well-defined objective evaluation metric that’s applied by a neutral third-party agency – in this case, the National Institute of Standards and Technologies.

It would be applied on shared datasets, and this would ensure that simple, clear knowledge is gained, because the participants have to reveal their methods to the sponsor and to one another at the time that the evaluation results are revealed.

And in 1986, in the United States, there couldn’t have been any other kind of automatic speech research program that could have gotten the funding.

So, in 1985, Dave Palat at NIST wrote a paper called “Performance Assessment of Automatic Speech Recognizers,” and I won’t read this at length, but the basic idea was that replicability is key. That you have to have published data; you have to have explicit, published, quantitative evaluation criteria. You put the data into the recognizer; you apply the evaluation program; you get out a number.

Have some other recognizer put in the same data, look at the output, compare it through the published evaluation criterion, which is generally published in the form of a program, as well as a description, you get out a number.

So, this resulted in what came to be called the “common task structure.” There was a detailed evaluation plan, often quite a long document that was developed in consultation with researchers and was published as the first step in the project. There was automatic evaluation software, producing a quantitative evaluation that was written and maintained by NIST. It was also published at the start of the project.

And, critically, there was shared data, training, and dev. test or development test data that was published at the start of the project, and then the eval. Test, evaluation test data would be withheld for periodic public evaluations.

Now, not everybody liked this, to say the least. A lot of the Piercians were skeptical. The basic idea was, “It doesn’t matter what you measure, you can’t turn water into gasoline.” And a lot of researchers were disgruntled. One well-known researcher in this area told me at the time, “It’s like being in first grade again. You’re told exactly what to do, and then you’re tested over and
over.” You know, it’s like “No Child Left Behind” for speech technology research.

But it worked. Why? The most obvious reason was that it allowed funding to start; it turned the spigot on. It also allowed funding to continue, because the funders could measure progress over time and point to their superiors at the graphs that showed that things were genuinely and believably getting better, even though the technology was still not ready for application.

Less obviously, it allowed project internal hill climbing, because the evaluation metrics were automatic, and the evaluation code was public, and so, the obvious way of working was to develop a new algorithm, or, more likely, a slight tweak on an existing algorithm, run it on the eval. test set, see if you get better. If you do, you keep the innovation. If it doesn’t, you throw it away.

So, you’re doing, in effect, algorithmic hill climbing, kind of like, you know, improving the internal combustion engine over the course of 50 years or so.

And the people who were complaining about being tested every six months actually started testing themselves several times a day in order to improve.

And perhaps the least obvious thing in advance, but the most obvious in retrospect, was that it created a culture because researchers shared methods and results, on shared data, with a common metric. And so, there was a—a—there came to be a large group of researchers, in the U.S. and around the world, who really spoke a common language because they had common problems and common solutions.

And in fact, participation in this culture became so valuable, that many research groups began to join without funding. And in fact, after a few years, DARPA realized that they could get research done in areas of interest to them without funding it, just by creating one of these common task evaluations and inviting people to come compete for gold stars. And people did.

It created a positive feedback loop, when everybody’s program has to interpret the same evidence, ambiguity resolution becomes a sort of gambling game, and this rewards the use of statistical methods. And given the nature of speech and language, statistical methods need the largest possible training set, which reinforces the value of
shared data, and these iterated train-and-test cycles allowed you to do the kind of algorithmic hill climbing that I talked about.

Now, over the past 25 years, this method has been applied to lots and lots of other problems: machine translation, speaker identification, language identification, parsing, sense disambiguation, information retrieval, information extraction, summarization, question answering, optical character recognition, sentiment analysis, and on and on.

The general experience is that error rates – as long as people keep working on the problem, error rates decline by a fixed percentage every year. That’s a relative percentage, obviously, not an absolute percentage, or you would soon get error rates below zero. But you get a kind of asymptotic decay to some error rate that’s determined basically by the noise and level of noise in the data.

Progress usually comes from many small improvements, and a change of one percent in performance can be a reason to break out the champagne. The conferences in this area are either heartening or depressing, depending on your attitude, because you hear talk after talk after talk in which somebody says, “Well, we did this incredibly elaborate experiment, and we improved performance by three-quarters of a percent.” And then there’s another talk, “And we did this other elaborate thing, and we improved performance by half a percent,” and so on.

And, you know, an improvement of one percent is a big deal. So, you could view that as being depressing. I mean where is the basic scientific advance? But of course, if you’ve got 50 improvement by half a percent, that adds up over time, as long as at least they’re semi independent and don’t cancel one another out.

The shared data plays a crucial role. It can be reused in unexpected ways, and glamour and deceit have been avoided, and a sort of self-sustaining process was started. So, I’m going to skip over a bunch of slides that basically show that if you do a Google Scholar search for some of the datasets that have been published, in association with these evaluations, you get counts like 12,000. This is 335 in the past year for that one, 6,000 and 192 in the past year, and so.

The number of tracks that the number of kinds of evaluation that NIST runs, or has run over the last 25 years, is numbered in the dozens. So, this happens to be one about text analysis. So, there’s question answering; recognizing textual entailment;
summarization; and what they call “knowledge base population,” which is sort of learning facts from bodies of text.

So, skipping over a bunch of this. The same kind of pattern now exists in lots of non-governmental entities. So, CoNLL – C-O-N-L-L – is the association for computing machineries. It’s a special interest group on natural language learning. At their annual meetings, since 1999, they’ve included a shared task in which training and test data is provided by the organizers, which allows participating systems to be evaluated and compared in a systematic way.

This is a technical society. This is not a government agency. Nobody’s – entirely volunteer; nobody’s paying for any – well, somehow it’s getting paid for, but people are doing it because they want to and not because someone is making them.

I’ll now give a small plug for my own organization, the Linguistic Data Consortium, which was kind of founded in the early days of this process by – with seed money from DARPA, to help create and maintain and distribute the data involved. And over our 20 years of existence, we’ve distributed more than 90,000 copies, of more than 1,300 titles, to more than 3,200 organizations, in 70-odd companies.

About half of the titles are common task datasets developed for these technology evaluation programs, and we add about 30 titles to our catalog every year.

Now, in 1983, the first conference on applied natural language processing had 34 presentations, none of which used a published dataset, none of which used a formal evaluation metric. That doesn’t mean they weren’t interesting. Here’s a couple of examples. Wendy Lehnert and Steven Schwartz, “Natural Language Processing System for Oil Exploration.” They described the problem in system architecture; they give examples of queries and responses.

Larry Reeker et al, “Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions.” So, this is an attempt to do information extraction from the chemical literature.

In 2010, there were 274 presentations at a comparable meeting, and essentially every single one of them used published data and published evaluation metrics from one or another of these
evaluation series. And the few that didn’t deal with a new dataset creation or new evaluation metrics. And here are a few examples, which I’ll skip over.

The point is that this is more or less the state of this field now. This is how people work in the area of human language technology exploration. That is, the research is almost inevitably on published sharable datasets using – by reference to published evaluation metrics, so that replication in reference to previous or subsequent algorithms is available.

There are some exceptions, especially now, because there are some companies like Google and Facebook and so on, Twitter, that have incredible bodies of data internally that they can’t publish for privacy reasons, or for reasons of confidentiality, or company advantage, corporate advantage.

And so, some of the people working in those areas do publish results on datasets that – where things can’t really be replicated. You sort of have to take their word for it. But that’s a minority; that remains a minority of a situation.

So, the culture of social science and the survey area is somewhat different, but maybe not all that different. I mean when I look over the history of the ANES survey, for example, which I’ve done recently, I’m certainly struck by the fact that many of the same values are kind of implicit in its design and it’s evolution.

Sharing data and problems lowers cost and barriers to entry. It creates intellectual communities, in which that body of data certainly has done. It speeds up replication and extension, and it guards against most forms of glamour and deceit, as well as simple confusion.

The thing that hasn’t quite happened, and someone this morning said that it would be helpful to some kinds of research, for people involved in survey work to loosen up a little bit with respect to sharing data with other researchers. And I would like to strongly underline and support that view. And I’d like to give – I know that there are serious problems about privacy and confidentiality and IRB protocols and so on, but I think that those can – those can be negotiated; those can be dealt with where there’s a will.

And the example that I’d like to give, as a point of reference, that I invite you to look into, is something called ADNI for the
Alzheimer’s Disease Neuroimaging Initiative. How many of you have heard of that? Anyone?

Okay, so this is something that was started by NIMH, National Institutes of Mental Health, and it enrolled – it has enrolled a very large number of patients funneled through, I think, at least 30 clinics.

For each of these patients, there’s detailed demographic information and medical histories. There’s FMRI scans, and I think structural MRI scans as well, in some cases PET scans. There’s blood tests. There’s cerebrospinal fluid, assays. There’s cognitive tests, and there’s physicians’ notes from the examination.

Obviously, the patients are anonymized, that is, you don’t know their name and address. They’re given some kind of key. But all of the rest of that information is, in effect, available to anyone. You could get it.

Now, it’s not up on the Internet for anybody to download. You have to send them a note and tell them what you want to do with it and sort of establish that you’re a bona fide researcher. And you have to sign something that says – imposes some protections on what you’re going to do with the data. But once you’ve done that, and it’s not hard to do, you can get all of this stuff.

Why did they do that? Because the current state of the art, as I understand it, what I’m told, is that if I were to go to see a neurologist today, exhibiting symptoms of “mild cognitive impairment,” as they call it, that is, my short-term memory is beginning to decay; my ability to remember names is even worse than it ever was, let’s say. Some other things I’m beginning to worry a little bit, so he does and examination.

He does a brain scan; he takes blood; he takes cerebrospinal fluid; he puts me through some cognitive tests, and he can tell me where I am. You know, what – he can put me in a diagnostic space. He can tell me nothing about what to expect. He can tell me, you know, “In five years, you could be a vegetable. In five years you could be more or less just like you are, maybe a little bit worse.” There’s nothing in that body of data that enables people to do a reliable projection any significant distance into the future.

So, what they would like to do is to encourage all sorts of people: computer scientists, statisticians, other biomedical researchers, anybody with any reasonable standing to take a flying leap at this
task, to come in and see if they can do better. Because they’ve got longitudinal data from these patients, and more accumulating all the time.

Now, if they can do that with complete medical records for thousands of people, I think it ought to be possible to do something similar with opinion survey results, or lifestyle survey results for similarly sized collections.

Obviously it requires appropriate IRB protocols to be filed. It requires appropriate informed consent on the part of the participants. It requires some kind of procedure for sort of taking care of how the data is handed out and what’s done with it after it’s handed out. But it seems to me that it should both legally, and from the point of view of government policy, and also ethically be quite feasible.

Okay, thank you. We have about 15 minutes for questions, I guess, if we extend the period. Yes?

**Audience:** In terms of the process, the training dataset is always troublesome to create and costs a lot of money.

**Mark Liberman:** Absolutely, yes.

**Audience:** Is there research in this area bout the process of how you create the trained dataset that can reduce that? So, you know, if you have multiple variables when you make the dataset – or the way I think about the training set issue, here’s the answer, and here’s the data, and this is – so, that’s what you train on. Is there anything that you can do in the training part to include process variables, to reduce the size of the dataset that’s needed for training?

**Mark Liberman:** Well, there are several strands – there’s probably less research on this than there deserves to be, frankly, because the costs involved in creating adequate training data in these areas is quite large, as you say, and the ability to shrink the costs makes a big difference.

One is the idea that we heard about before lunch, better tools and semi-automated annotation, where you introduce some programs that maybe do a first draft that you check, or at least enable you to negotiate the annotation process in fewer seconds per bite, so to speak, per data item.

That sort of thing is extremely important, and it can make a big difference. A well-designed tool can easily decrease the labor time
by a factor of two or so, and if you add in semi-automated filling in of the things that the computer can do well, that can multiply by another factor.

A second thing that happens, that I think has been a benefit of these large organized DARPA programs is that the data is fed in increments to the researchers as part of the program, and the researchers (a) always want more for the same amount of money, but (b) they also always want the quality to be higher, and they’re always checking on stuff.

And they’re saying, “Look, here you said this, and there they said that. What’s going on here,” kind of thing, “and by the way, we don’t need anymore of this stuff because we’re already doing well at that, but we do need more of that stuff.” So, in general, datasets that are sort of created and being used at the same time are often better for those reasons, or at least more cost effective.

And then finally, there are some ideas, like so-called active learning, which you’ve probably heard of. The idea there is that you have some training data, and you use that. Now you have a large amount of unannotated data, and you use what you’ve learned in the training to try to figure out which pieces of the unannotated data you would learn the most from getting annotated.

In effect, that’s trying to automate this observation that I attributed to a researcher earlier, which is, “We’re already doing really well on this part, so we don’t need anymore of those, but we hardly have any samples of this kind, so let’s get some more of those, so we can see how they all – how it all shakes out.”

And that – you know, if you were to go to Google Scholar and look for active learning, you would find thousands and thousands of articles. And unfortunately, about half of them would be cases in which it’s shown that the active learning experiment didn’t work any better than random selection of things to annotate, but half of them would be cases where they claimed an advantage. So, that’s certainly one area.

Yeah?

*Audience:* This is a question about where this machine things has worked most effectively and less effectively. We now disambiguate names, and there’s a bunch of other information that goes in to determine who wrote scholarly papers and whether it’s the same person.
That seems to be fairly well developed. Then you get to abstracts of papers, or actually determining what’s in a paper, and then we have to get the sentiments, be it political sentiment on Twitter or something. Could you give a little specificity to those areas?

Mark Liberman: Well, of course, a big part of this has to do with how crisp the answers are, so to speak. So, with something like figuring out who wrote which paper, reasonably well-trained people, who – you know, librarians, for example, but even ordinary people with an understanding of how these things work, can figure out and agree on the answer to those questions, that is, “Are these two – you know, is J. F. Smith the same as John F. Smith in this other paper,” with a high degree of agreement.

And that’s a big help, but that doesn’t solve the problem. But at least it means you can get data that’s almost – you know, that has a very low – a very higher inter-annotator agreement rate.

At the other end of the spectrum, things like emotion detection in speech, for example, there are lots of schemes for coding emotion, and it’s very hard to find any of them where independent annotators – there’s information there, there’s no question about it, but the number of bits per judgment effectively is not all that high.

So, that’s either good news or bad news, depending on your opinion. The good news is that it’s easy to have – to write a program that agrees with human annotators, as well as they agree with each other. But the bad news is that they don’t agree with each other very well.

And unfortunately, that isn’t always a very good metric, because it’s sometimes the case that the machine makes very different kinds of mistakes, or very different kinds of deviations than the humans do. This is the sort of thing that has come up, for example, in the Educational Testing Service’s attempts to develop its e-rater program, which automatically grades SAT essays and other kinds of essays. And that actually does quite well in the sense that the grades that come out of that agree with human judges about as well as they agree with one another.

Again, the one piece of the bad news is that they don’t agree with one another all that well, unfortunately, and the other piece of bad news is that because of the way e-rater works, you could – if you knew how it worked, you could game the system by producing an
essay that would be completely incoherent gibberish, and would actually score quite high.

Now, of course, the students who take the SAT don’t – aren’t clever enough to do that, and then why would they want to anyway? But it gives you the sense that there’s something still missing from the procedure.

So, the way they use it is rather like the way we heard about the answer classifications being used in the ANES project, namely they have – they used to have each essay graded by two humans, and if they agreed close enough, then that was accepted, and if they disagreed, then a senior judge would come along and adjudicate. Now they have one human and the machine, and if they agree close enough, then that’s accepted, and if they don’t, then another human comes along and adjudicates. So, that cuts out half their labor cost, basically, and they’re very happy about that.

Audience: And who was that you just described in that procedure?

Mark Liberman: Education Testing Service.

Audience: Oh, they did that, too?

Mark Liberman: Yeah, that’s how they do it now, as I understand it. And they’re working hard to develop similar kinds of approaches for doing TOEFL verbal facility ratings and stuff like that. So, I think it’s very difficult to give an absolute answer about how well different things are going to work.

But – because it depends, as I said, on the quality and consistency of the data. It depends on kind of the intrinsic difficulty of the task. It depends on the fit to the task of the particular model that’s used. And all of those things tend to get better as people work on the problem over time.

So, they try out lots of different models, and keep the ones that work better. They work on making the sort of annotation guidelines or classification guidelines clearer, as the ANES people did, and more consistent. And they just sort of generally work on trying to get things to work better rather than worse. And if – you know, I think that sort of human observation is that in pretty much any area – there’s some where it doesn’t work – but in almost any area, where you can at least get that process started, if people keep after it for a long time, they get better as a community.
Yeah?

_Audience:_ I know Phil Stone – I don’t know if you know, Phil Stone is working on this for verbatim reports for surveys, and for language disambiguation to determine whether the people’s responses to the verbatim were generally more positive, more negative, and so forth.

And it seems like there’s a lot of overlap there. But much of what we do in survey research has a very, very limited range of responses that are necessary. It seems to me that the survey researchers have been very, very slow to adopt these. And the reason I say that is I used to fly a lot with United Airlines, and they used to lose my bags rather routinely. And so, I got to know Simon very well.

I don’t know if any of you fly with United, but you would call, and it was very clearly that you were speaking to a computer. But it was a modulated, very pleasant voice. And since I had a lot of opportunities to talk with Simon, I started to play around, and I realized that, he’s like, “Most of the people who call me, call because they’ve lost a bag, how about you?”

“Yeah.”

[Laughter]

_Audience:_ Or, “Sure,” or, “Yes.” There were a wide range of – but, of course, I’m viewing this as a wide range, but it’s a very limited range of language that’s being used, and that’s true of most of what we do in survey research.

_Mark Liberman:_ Well, it’s also – it’s certainly something that people aim at when they’re designing call center applications of that kind, that you want to put the user in a situation where the range of things that they’re likely to say is as small as possible, so that you have the best shot at deciding which one of them they meant.

_Audience:_ And so, kind of a follow-up on that is, we also know that when, on the telephone, that people – the respondents will guess at the age range, the race is a big issues. In fact, often on surveys, at the end, Gallup does an inventory, a social inventory survey each year on race, and they ask at the end of the survey – they ask the respondent to guess the race of the interviewer. And they go back to compare that to see – and they find the same patterns, but computer voices can be modulated.
And it seems to me to be an area where there could be some investigation, to modulate the voice as an African-American voice, as a woman, as a man, a younger person, an older person. This research that we have with humans, that could be useful to investigate to see, as we modulate these voices, would it be possible to match a voice so that it is less – if it’s an African-American respondent, to modulate the voice of the interviewer.

**Mark Liberman:** Yes, all of those things are, I think, doable. What you said, in addition, made me think of two other research projects that look interesting to me from the human language technology research side, but might also be interesting to you folks.

So, one of the things that we heard before lunch is that figuring out what categories to divide answers into is often unclear, sort of problematic. And someone raised the possibility of using focus group discussions, or something of that kind, or at least longer stretches of answer from a smaller number of people as a first step to try to sort of figure out what’s going on there.

There’s all kinds of interesting ideas about how to do clustering of answers in a body of material like that. And I think it would be very – so, in effect, what you’re doing, if you do that, as a human is that you’re reading over or listening to all of the focus group results, and you’re saying, “Oh, yes. Well, there are a lot of people who say this, and there are a lot of people who say that.” It might be helpful to have the results of a clustering analysis of the transcripts to look at when you made that determination.

The other thing that has occurred to me several times, and what you said reminded me of, is that as I said before, it’s really problematic to try to get human beings to judge sentiment or emotion or opinion and other things like that. The number of bits per judgment that you get out is just not that high, and the level of agreement is not that great.

But when you have a relatively large survey instrument that’s been applied to a large number of speakers, there’s an awful lot of implicit information in there that ought – you know, people’s answers to one question are obviously not uncorrelated with their answers to another question, and their intensity of feeling about one thing is not uncorrelated to their intensity and feeling about other things.
And if you had longitudinal surveys, in which you look at the ways in which people’s responses or attitudes change, then that time dimension adds some additional constraints. And I think it would be very interesting to try to look for some kind of emergent dimensions of opinion [slash] emotions [slash] sentiment rather than trying to think in advance what they should be, look for ways to get them to emerge from some kind of appropriate factorization of the body of data from the survey.

**Audience:**
I think it’s fair – we had a meeting at the agriculture department last week, and we were reminded of the fact that one of the founders of our field, a guy named Rensis Likert, ran a survey operation in the agriculture department.

And they made their reputation by the way they designed questionnaires, which is extremely ironic if you know anything about Rensis Likert. He’s known for inventing a scale – a five-point scale –

**Mark Liberman:** The Likert scale, yeah.

**Audience:** – that everybody talks about as a Likert scale. But that operation that he ran only asked open questions. And they had a large group of coders, highly trained and well compensated, whose job it was to make sense of these responses to open questions given by farmers. And they had a huge dispute at the end of the second World War with the so-called posters about how you ought to frame questionnaires because the posters did them in these closed questions, in essentially what you might call a Likert scale.

I think it’s fair to say, although, you know, I may be going way out on a limb here, I think it’s fair to say that an operation like the agriculture department’s one, would not ever happen today, and that we have moved, as a field, far away from using open questions.

And part of that is driven by, I think, by a change in the technologies we use to gather data. Part of it is – and part of that is in relationship to the fact that these technologies are frequently now self-administered, respondents are asked to fill things out by themselves, and asking respondents to type long sentences into Web questionnaires, or write them down is something that we all think is fairly difficult.

So, I think the feel is, in general – I remember hearing Provost Grove say, at one point, “The CATI interview, the telephone
interview destroyed the open question.” I don’t buy this sort of technological determinism stuff, but in any case, there was a point to that.

So, now I think we hear about the notion of understanding open material, textual corpora, as my former colleagues used to call, in the terms of data mining, which is mining big, big corpora of texts, Twitter, whatever, and their whole business is now set up to compete with the survey business on the basis of analyzing these large corpora.

So, where am I going with this? I think where I’m going –

[Laughter]

Mark Liberman: For the future.

Audience: So, I think where I’m going is that, with the exception of something like the ANES, where they actually have – they come from a time when the open question was quite common, and they have to replicate it. Today, we’re moving more and more away from the open question.

And so, maybe an opportunity for us, I mean apart from the ones Alan has noted, and these cases where you’ve been speaking about today, and opportunity for us is to find ways of melding the material that is not survey based with the stuff that is survey based, to complement what we know. Because I think surveys are not designed to capture those data as well as they used to.

Mark Liberman: If I – just 30 seconds. I would just like to note that I’m very interested in the idea that you could learn some things from the way people verbally answer closed questions. So, if rather than having the interviewer quoted, you just recorded what they say when you ask them, “Who are you going to vote for? What’s your registration? You know, what party are you registered with?” I think there would be a lot of extra information in exactly what they said and how they said it.

Jon Krosnick: Okay, so I can’t resist editorializing a little bit. Tomorrow morning, Gary Langer, among other things, is going to talk about what Peter just raised, which is the idea of blending in large bodies of text with survey data and analysis.

And a second thing I’ll editorialize on, as Peter said, that it’s hard to imagine people giving – typing long answers. Open questions
may be dying in automated surveys. Actually, I think people who live in that world think the opposite, that interviewers abbreviate answers much too much.

We tried really hard recently with field interviewers to get them to tell the respondent to slow down, type every word, and don’t abbreviate at all. And then after the interview, if you had to abbreviate anything, go back and edit it so it’s complete sentences. Even after all that, we can’t get complete sentences. Whereas in the self-administered stuff, the respondents type out more. So, I’m not sure that technology’s going to kill that exactly.

But so let me keep us on schedule. So, I’ll welcome Colm to start transitioning here, while Alan makes his last comment here.

Allan McCutcheon: I don’t think people are going to be typing this in, and quite frankly, I don’t think it’s going to be typed in by the interviewer on CATI systems, almost everyone is now digitized, recording everything, and that can be transcribed — machine transcribed, not typed in by human beings. And I don’t think on Internet survey, frankly, we’ll ask people to type things in. We’ll ask them to speak — right? “Tell us what you think about this.” It gets recorded. There’s no reason to ask them to type it in. It reduces the respondent burden rather dramatically, and it can all be digitized; it’s digitally recorded, and it can be transcribed.

Jon Krosnick: So, this is a great idea in principle. In practice, Dragon and other software that are designed to — or Siri. I mean if you use Siri, it gets it some of the time after training. But when you’re recording survey response —

Allan McCutcheon: Well, we can’t plan for the technology of today. We’ve got to start thinking about the technology of tomorrow. Right? And the technology — I mean if you told people ten years ago that you’d be talking to your cell phone, they would have said, “Well, yeah, to my mother.” Right?

But, “No, no, no, you’ll be talking to your cell phone.” Right? They’d look at you as if you just had your head screwed on wrong. But today, people are doing it, and they’re saying, “Well, it doesn’t do it perfectly. It doesn’t understand me as well as my mother does.” Wait for ten years.