GoogleTalk

This topic has three different inter-related components; any of them can be selected–or you can extend it on your own.

Topic 1: You have just been hired as a consultant at Google, to assist with their new "GoogleTalk" project. Much like their Google Books project, GoogleTalk aims to collect many, many millions of examples of spoken sentences, initially just in English, transcribed into written text.
You overhear Google employee #25 talking to Google employee #200 about their plans for this data:

"Look," she says, "plainly, the number of sentences one person will *ever* hear or speak in a lifetime is finite – and so therefore is the collection of all the sentences we'll ever put into Google Speech, even if it's trillions and trillions of examples. This collection, a 'corpus,' constitutes the set of 'observables' for natural language. It is this corpus that we have to model. It's just like when we observe some other natural phenomenon, like the motion of the planets. We can use lots and lots of astronomical observations, and then, once we know the position of, say, Saturn at many points in time, we can predict where it will be in the next moment, just by using our collected data. So for example, the probability of where Saturn is at time *t* is just contingent on where it was at some finite measured number of instances in the past. With sentences, we can do the same thing via the method called *ngrams*. An *ngram* is just a way of predicting what the n^{th} word in a sentence will be, given the *n*–1 preceding words. And that's what we have lots of data about. We can use the probabilities of such sequences to capture what we need to know. For example, if we see the sequence "I'd like to make a collect…" then a very likely next word is *call*, or *phone*, or *international*, but not *the.* It should be a snap. We can do fancier statistics if we need to – I know that sometimes specific, very long sequences won't ever show up in our corpus, so they'll have a frequency of zero, but we now have sophisticated ways of estimating this kind of missing data."

Employee #200 replies, "Wait a minute. Are you sure that's the right thing to study? Isn't the set of sentences that even one person can potentially produce countably infinite? How do you determine what goes on your list, and what does not? And I'm a bit troubled by your physics analogy. I don't think Newton would have appreciated it. Sure, Copernicus and Kepler collected lots and lots of data sequences, but what underlies them, *F=ma*, isn't just a statistical approximation – it's an absolute principle. A *theory.* What you want to model – the true 'observables' – isn't what's in the 'outside world', the sequences of words or sentences, but rather the principles of the 'inside world' – the 'cognitive machinery' that produces or perceives this or that collection of sentences."

Keep these arguments in mind. You're going to write a brief report on them in a bit, but first you get some experience of your own with *n*-grams.

Now consider again the arguments by the Google employees: take the first employee's side and argue for her viewpoint. Then, switch sides to the second employee and argue **against** the first employee's viewpoint. A good consultant should be able to do a good job arguing both the positive and negative sides of a thesis. And it's quite common for one or both of the positions to be poorly formulated –people are only human. A good consultant should aim at a sympathetic interpretation no matter what the position and attempt to sharpen the proposals, since the goal isn't to score debating points with cheap, shallow, and easy shots.

Topic 2: Read the pdf extract by Chomsky associated with this topic, starting at section 36.1 from Chomsky's 1954-55 book on one difference between "probable" and "grammatical" and how statistics and linguistics might connect. This is the first appearance in print of the famous sentence

"Colorless green ideas sleep furiously," but unfortunately it's not often read in its full context. Pay attention to the empirical evidence that Chomsky uses to determine how we know that the example "colorless green ideas sleep furiously" is actually just as grammatical as revolutionary new ideas appear infrequently," even though their strict probability of occurrence might differ wildly. (As it happens, this approach was 're-discovered' 50 years later as an "exciting new development"–without the realization that it was a re-discovery.) Incidentally, Chomsky also mentions Markov models when he uses the (then conventional) term 'n^{th} order of approximation.' Note: this is *not* a straightforward debate–people are still arguing about it, and it speaks to the very essence of what it means to construct a scientific theory of human language.

.

Topic 3: Consider the following quote attributed to the late Fred Jelinek (1988): "*Anytime a linguist leaves the group our machine translation accuracy goes up.*" (Specifically, compare and contrast the following analogical aphorism in relationship to Jelinek's remark, in light of your reading and analysis above: "*Anytime a particle moving near the speed of light leaves the group, our classical mechanics prediction accuracy goes up."*) Discuss how this bears on the proper data to be used for the scientific analysis of human language.