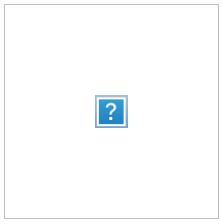
On Chomsky and the Two Cultures of Statistical Learning

At the Brains, Minds, and Machines symposium held during MIT's 150th birthday party, Technology Review reports that Prof. Noam Chomsky

> derided researchers in machine learning who use purely statistical methods to produce behavior that mimics something in the world, but who don't try to understand the meaning of that behavior.



MIT: 150

The transcript is now available, so let's quote Chomsky himself:

> It's true there's been a lot of work on trying to apply statistical models to various linguistic problems. I think there have been some successes, but a lot of failures. There is a notion of success ... which I think is novel in the history of science. It interprets success as approximating unanalyzed data.

This essay discusses what Chomsky said, speculates on what he might have meant, and tries to determine the truth and importance of his claims.

Chomsky's remarks were in response to Steven Pinker's question about the success of probabilistic models trained with statistical methods.

Noam Chomsky

1. What did Chomsky mean, and is he right?
2. What is a statistical model?
3. How successful are statistical language models?
4. Is there anything like their notion of success in the history of science?
5. What doesn't Chomsky like about statistical models?

What did Chomsky mean, and is he right?

I take Chomsky's points to be the following:

A. Statistical language models have had engineering success, but that is irrelevant to science.
B. Accurately modeling linguistic facts is just butterfly collecting; what matters in science (and specifically linguistics) is the underlying principles.
C. Statistical models are incomprehensible; they provide no insight.
D. Statistical models may provide an accurate simulation of some phenomena, but the simulation is done completely the wrong way; people don't decide what the third word of a sentence should be by consulting a probability table keyed on the previous two words, rather they map from an internal semantic form to a syntactic tree-structure, which is then linearized into words. This is done without any probability or statistics.
E. Statistical models have been proven incapable of learning language; therefore language must be innate, so why are these statistical modelers wasting their time on the wrong enterprise?

Is he right? That's a long-standing debate. These are my answers:

A. I agree that engineering success is not the goal or the measure of science. But I observe that science and engineering develop together, and that engineering success shows that something is working right, and so is evidence (but not proof) of a scientifically successful model.
B. Science is a combination of gathering facts and making theories; neither can progress on its own. I think Chomsky is wrong to push the needle so far towards theory over facts; in the history of science, the laborious accumulation of facts is the dominant mode, not a novelty. The science of understanding language is no different than other sciences in this respect.

C. I agree that it can be difficult to make sense of a model containing billions of parameters. Certainly a human can't understand such a model by inspecting the values of each parameter individually. But one can gain insight by examing the *properties* of the model—where it succeeds and fails, how well it learns as a function of data, etc.
D. I agree that a Markov model of word probabilities cannot model all of language. It is equally true that a concise tree-structure model without probabilities cannot model all of language. What is needed is a probabilistic model that covers words, trees, semantics, context, discourse, etc. Chomsky dismisses all probabilistic models because of shortcomings of particular 50-year old models. I understand how Chomsky arrives at the conclusion that probabilistic models are unnecessary, from his study of the generation of language. But the vast majority of people who study *interpretation* tasks, such as speech recognition, quickly see that interpretation is an inherently probabilistic problem: given a stream of noisy input to my ears, what did the speaker most likely mean? Einstein said to make everything as simple as possible, but no simpler. Many phenomena in science are stochastic, and the simplest model of them is a probabilistic model; I believe language is such a phenomenon and therefore that probabilistic models are our best tool for representing facts about language, for algorithmically processing language, and for understanding how humans process language.
E. In 1967, Gold's Theorem showed some theoretical limitations of logical deduction on formal mathematical languages. But this result has nothing to do with the task faced by learners of natural language. In any event, by 1969 we knew that probabilistic inference (over probabilistic context-free grammars) is not subject to those limitations (Horning showed that learning of PCFGs is possible). I agree with Chomsky that it is undeniable that humans have some innate capability to learn natural language, but we don't know enough about that capability to rule out probabilistic language representations, nor statistical learning. I think it is much more likely that human language learning involves something like probabilistic and statistical inference, but we just don't know yet.

Now let me back up my answers with a more detailed look at the remaining questions.

What is a statistical model?

A **statistical model** is a mathematical model which is modified or trained by the input of data points. Statistical models are often but not always probabilistic. Where the distinction is important we will be careful not to just say "statistical" but to use the following component terms:

- A **mathematical model** specifies a relation among variables, either in functional form that maps inputs to outputs (e.g. $y = m\,x + b$) or in relation form (e.g. the following (x, y) pairs are part of the relation).
- A **probabilistic model** specifies a probability distribution over possible values of random variables, e.g., $\mathrm{P}(x, y)$, rather than a strict deterministic relationship, e.g., $y = f(x)$.
- A **trained model** uses some training/learning algorithm to take as input a collection of possible models and a collection of data points (e.g. (x, y) pairs) and select the best model. Often this is in the form of choosing the values of parameters (such as m and b above) through a process of statistical inference.


Claude Shannon

For example, a decade before Chomsky, Claude Shannon proposed probabilistic models of communication based on Markov chains of words. If you have a vocabulary of 100,000 words and a second-order Markov model in which the probability of a word depends on the previous two words, then you need a quadrillion (10^{15}) probability values to specify the model. The only feasible way to learn these 10^{15} values is to gather statistics from data and introduce some smoothing method for the many cases where there is no data. Therefore, most (but not all) probabilistic models are trained. Also, many (but not all) trained models are probabilistic.

As another example, consider the Newtonian model of gravitational attraction, which says that the force between two objects of mass m_1 and m_2 a distance r apart is given by

$$F = G\, m_1\, m_2 / r^2$$

where G is the universal gravitational constant. This is a trained model because the gravitational constant G is determined by statistical inference over the results of a series of experiments that contain stochastic experimental error. It is also a deterministic (non-probabilistic) model because it states an exact functional relationship. I believe that Chomsky has no objection to this kind of statistical model. Rather, he seems to reserve his criticism for statistical models like Shannon's that have quadrillions of parameters, not just one or two.

(This example brings up another distinction: the gravitational model is **continuous** and **quantitative** whereas the linguistic tradition has favored models that are **discrete**, **categorical**, and **qualitative**: a word is or is not a verb, there is no question of its degree of verbiness. For more on these distinctions, see Chris Manning's article on Probabilistic Syntax.)

A relevant probabilistic statistical model is the ideal gas law, which describes the pressure P of a gas in terms of the the number of molecules, N, the temperature T, and Boltzmann's constant, K:

$$P = N\, k\, T / V.$$

The equation can be derived from first principles using the tools of statistical mechanics. It is an uncertain, incorrect model; the *true* model would have to describe the motions of individual gas molecules. This model ignores that complexity and *summarizes* our uncertainty about the location of individual molecules. Thus, even though it is statistical and probabilistic, even though it does not completely model reality, it does provide both good predictions and insight—insight that is not available from trying to understand the *true* movements of individual molecules.

Now let's consider the non-statistical model of spelling expressed by the rule "*I before E except after C*." Compare that to the probabilistic, trained statistical model:

```
P(IE) = 0.0177          P(CIE) = 0.0014          P(*IE) = 0.163
P(EI) = 0.0046          P(CEI) = 0.0005          P(*EI) = 0.0041
```

This model comes from statistics on a corpus of a trillion words of English text. The notation P(IE) is the probability that a word sampled from this corpus contains the consecutive letters "IE." P(CIE) is the probability that a word contains the consecutive letters "CIE", and P(*IE) is the probability of any letter other than C followed by IE. The statistical data confirms that IE is in fact more common than EI, and that the dominance of IE lessens wehn following a C, but contrary to the rule, CIE is still more common than CEI. Examples of "CIE" words include "science," "society," "ancient" and "species." The disadvantage of the "I before E except after C" model is that it is not very accurate. Consider:

```
Accuracy("I before E") = 0.0177/(0.0177+0.0046) = 0.793
Accuracy("I before E except after C") = (0.0005+0.0163)/(0.0005+0.0163+0.0014+0.0041) = 0.753
```

A more complex statistical model (say, one that gave the probability of all 4-letter sequences, and/or of all known words) could be ten times more accurate at the task of spelling, but offers little **insight** into what is going on. (Insight would require a model that knows about phonemes, syllabification, and language of origin. Such a model could be trained (or not) and probabilistic (or not).)

As a final example (not of statistical models, but of insight), consider the Theory of Supreme Court Justice Hand-Shaking: when the supreme court convenes, all attending justices shake hands with every other justice. The number of attendees, n, must be an integer in the range 0 to 9; what is the total number of handshakes, h for a given n? Here are three possible explanations:

A. Each of n justices shakes hands with the other n - 1 justices, but that counts Alito/Breyer and Breyer/Alito as two separate shakes, so we should cut the total in half, and we end up with $h = n \times (n - 1) / 2$.

B. To avoid double-counting, we will order the justices by seniority and only count a more-senior/more-junior handshake, not a more-junior/more-senior one. So we count, for each justice, the shakes with the more junior justices, and sum them up, giving $h = \Sigma_{i = 1 .. n} (i - 1)$.

C. Just look at this table:

n:	0	1	2	3	4	5	6	7	8	9
h:	0	0	1	3	6	10	15	21	28	36

Some people might prefer A, some might prefer B, and if you are slow at doing multiplication or addition you might prefer C. Why? All three explanations describe *exactly the same theory* — the same function from n to h, over the entire domain of possible values of n. Thus we could prefer A (or B) over C only for reasons other than the theory itself. We might find that A or B gave us a better understanding of the problem. A and B are certainly more useful than C for figuring out what happens if Congress exercises its power to add an additional associate justice. Theory A might be most helpful in developing a theory of handshakes at the end of a hockey game (when each player shakes hands with players on the opposing team) or in proving that the number of people who shook an odd number of hands at the MIT Symposium is even.

How successful are statistical language models?

Chomsky said words to the effect that statistical language models have had some limited success in some application areas. Let's look at computer systems that deal with language, and at the notion of "success" defined by "making accurate predictions about the world." First, the major application areas:

- **Search engines:** 100% of major players are trained and probabilistic. Their operation cannot be described by a simple function.
- **Speech recognition:** 100% of major systems are trained and probabilistic, mostly relying on probabilistic hidden Markov models.
- **Machine translation:** 100% of top competitors in competitions such as NIST use statistical methods. Some commercial systems use a hybrid of trained and rule-based approaches. Of the 4000 language pairs covered by machine translation systems, a statistical system is by far the best for every pair except Japanese-English, where the top statistical system is roughly equal to the top hybrid system.
- **Question answering:** this application is less well-developed, and many systems build heavily on the statistical and probabilistic approach used by search engines. The IBM Watson system that recently won on Jeopardy is thoroughly probabilistic and trained, while Boris Katz's START is a hybrid. All systems use at least some statistical techniques.

Now let's look at some components that are of interest only to the computational linguist, not to the end user:

- **Word sense disambiguation:** 100% of top competitors at the SemEval-2 competition used statistical techniques; most are probabilistic; some use a hybrid approach incorporating rules from sources such as Wordnet.
- **Coreference resolution:** The majority of current systems are statistical, although we should mention the system of Haghighi and Klein, which can be described as a hybrid system that is mostly rule-based rather than trained, and performs on par with top statistical systems.
- **Part of speech tagging:** Most current systems are statistical. The Brill tagger stands out as a successful hybrid system: it learns a set of deterministic rules from statistical data.
- **Parsing:** There are many parsing systems, using multiple approaches. Almost all of the most successful are statistical, and the majority are probabilistic (with a substantial minority of deterministic parsers).

Clearly, it is inaccurate to say that statistical models (and probabilistic models) have achieved *limited* success; rather they have achieved a *dominant* (although not exclusive) position.

Another measure of success is the degree to which an idea captures a community of researchers. As Steve Abney wrote in 1996, "In the space of the last ten years, statistical methods have gone from being virtually unknown in computational linguistics to being a fundamental given. ... anyone who cannot at least use the terminology persuasively risks being mistaken for kitchen help at the ACL [Association for Computational Linguistics] banquet."

Now of course, the majority doesn't rule -- just because everyone is jumping on some bandwagon, that doesn't make it right. But I made the switch: after about 14 years of trying to get language models to work using logical rules, I started to adopt probabilistic approaches (thanks to pioneers like Gene Charniak (and Judea Pearl for probability in general) and to my colleagues who were early adopters, like Dekai Wu). And I saw everyone around me making the same switch. (And I didn't see anyone going in the other direction.) We all saw the limitations of the old tools, and the benefits of the new.

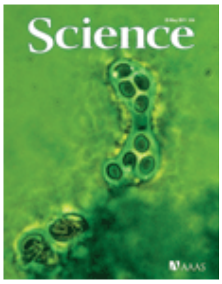
And while it may seem crass and anti-intellectual to consider a financial measure of success, it is worth noting that the intellectual offspring of Shannon's theory create several trillion dollars of revenue each year, while the offspring of Chomsky's theories generate well under a billion.

This section has shown that one reason why the vast majority of researchers in computational linguistics use statistical models is an *engineering* reason: statistical models have state-of-the-art performance, and in most cases non-statistical models perform worst. For the remainder of this essay we will concentrate on *scientific* reasons: that probabilistic models better represent linguistic facts, and statistical techniques make it easier for us to make sense of those facts.

Is there anything like [the statistical model] notion of success in the history of science?

When Chomsky said "*That's a notion of [scientific] success that's very novel. I don't know of anything like it in the history of science*" he apparently meant that the notion of success of "accurately modeling the world" is novel, and that the only true measure of success in the history of science is "providing insight" — of answering *why* things are the way they are, not just describing *how* they are.

A dictionary definition of science is "the systematic study of the structure and behavior of the physical and natural world through observation and experiment," which stresses accurate modeling over insight, but it seems to me that both notions have always coexisted as part of doing science. To test that, I consulted the epitome of doing science, namely *Science*. I looked at the current issue and chose a title and abstract at random:

> Chlorinated Indium Tin Oxide Electrodes with High Work Function for Organic Device Compatibility
>
> 
>
>
> In organic light-emitting diodes (OLEDs), a stack of multiple organic layers facilitates charge flow from the low work function [~4.7 electron volts (eV)] of the transparent electrode (tin-doped indium oxide, ITO) to the deep energy levels (~6 eV) of the active light-emitting organic materials. We demonstrate a chlorinated ITO transparent electrode with a work function of >6.1 eV that provides a direct match to the energy levels of the active light-emitting materials in state-of-the art OLEDs. A highly simplified green OLED with a maximum external quantum efficiency (EQE) of 54% and power efficiency of 230 lumens per watt using outcoupling enhancement was demonstrated, as were EQE of 50% and power efficiency of 110 lumens per watt at 10,000 candelas per square meter.

It certainly seems that this article is much more focused on "accurately modeling the world" than on "providing insight." The paper does indeed fit in to a body of theories, but it is mostly reporting on specific experiments and the results obtained from them (e.g. efficiency of 54%).

I then looked at all the titles and abstracts from the current issue of *Science*:

- Comparative Functional Genomics of the Fission Yeasts
- Dimensionality Control of Electronic Phase Transitions in Nickel-Oxide Superlattices
- Competition of Superconducting Phenomena and Kondo Screening at the Nanoscale
- Chlorinated Indium Tin Oxide Electrodes with High Work Function for Organic Device Compatibility
- Probing Asthenospheric Density, Temperature, and Elastic Moduli Below the Western United States
- Impact of Polar Ozone Depletion on Subtropical Precipitation
- Fossil Evidence on Origin of the Mammalian Brain
- Industrial Melanism in British Peppered Moths Has a Singular and Recent Mutational Origin
- The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants
- Chromatin "Prepattern" and Histone Modifiers in a Fate Choice for Liver and Pancreas
- Spatial Coupling of mTOR and Autophagy Augments Secretory Phenotypes
- Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans
- The Toll-Like Receptor 2 Pathway Establishes Colonization by a Commensal of the Human Microbiota
- A Packing Mechanism for Nucleosome Organization Reconstituted Across a Eukaryotic Genome
- Structures of the Bacterial Ribosome in Classical and Hybrid States of tRNA Binding

and did the same for the current issue of *Cell*:

- Mapping the NPHP-JBTS-MKS Protein Network Reveals Ciliopathy Disease Genes and Pathways
- Double-Strand Break Repair-Independent Role for BRCA2 in Blocking Stalled Replication Fork Degradation by MRE11
- Establishment and Maintenance of Alternative Chromatin States at a Multicopy Gene Locus
- An Epigenetic Signature for Monoallelic Olfactory Receptor Expression
- Distinct p53 Transcriptional Programs Dictate Acute DNA-Damage Responses and Tumor Suppression
- An ADIOL-ERβ-CtBP Transrepression Pathway Negatively Regulates Microglia-Mediated Inflammation
- A Hormone-Dependent Module Regulating Energy Balance
- Class IIa Histone Deacetylases Are Hormone-Activated Regulators of FOXO and Mammalian Glucose Homeostasis

and for the 2010 Nobel Prizes in science:

- Physics: *for groundbreaking experiments regarding the two-dimensional material graphene*
- Chemistry: *for palladium-catalyzed cross couplings in organic synthesis*
- Physiology or Medicine: *for the development of in vitro fertilization*

My conclusion is that 100% of these articles and awards are more about "accurately modeling the world" than they are about "providing insight," although they all have some theoretical insight component as well. I recognize that judging one way or the other is a difficult ill-defined task, and that you shouldn't accept my judgements, because I have an inherent bias. (I was considering running an experiment on Mechanical Turk to get an unbiased answer, but those familiar with Mechanical Turk told me these questions are probably too hard. So you the reader can do your own experiment and see if you agree.)

What doesn't Chomsky like about statistical models?

I said that statistical models are sometimes confused with probabilistic models; let's first consider the extent to which Chomsky's objections are actually about probabilistic models. In 1969 he famously wrote:

> But it must be recognized that the notion of "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

His main argument being that, under any interpretation known to him, the probability of a novel sentence must be zero, and since novel sentences are in fact generated all the time, there is a contradiction. The resolution of this contradiction is of course that it is not necessary to assign a probability of zero to a novel sentence; in fact, with current probabilistic models it is well-known how to assign a non-zero probability to novel occurrences, so this criticism is invalid, but was very influential for decades. Previously, in *Syntactic Structures* (1957) Chomsky wrote:

> I think we are forced to conclude that ... probabilistic models give no particular insight into some of the basic problems of syntactic structure.

In the footnote to this conclusion he considers the possibility of a useful probabilistic/statistical model, saying "I would certainly not care to argue that ... is unthinkable, but I know of no suggestion to this effect that does not have obvious flaws." The main "obvious flaw" is this: Consider:

1. **I** never, ever, ever, ever, ... **fiddle** around in any way with electrical equipment.
2. **She** never, ever, ever, ever, ... **fiddles** around in any way with electrical equipment.
3. * **I** never, ever, ever, ever, ... **fiddles** around in any way with electrical equipment.
4. * **She** never, ever, ever, ever, ... **fiddle** around in any way with electrical equipment.

No matter how many repetitions of "ever" you insert, sentences 1 and 2 are grammatical and 3 and 4 are ungrammatical. A probabilistic Markov-chain model with *n* states can never make the necessary distinction (between 1 or 2 versus 3 or 4) when there are more than *n* copies of "ever." Therefore, a probabilistic Markov-chain model cannot handle all of English.

This criticism is correct, but it is a criticism of Markov-chain models—it has nothing to do with probabilistic models (or trained models) at all. Moreover, since 1957 we have seen many types of probabilistic language models beyond the Markov-chain word models. Examples 1-4 above can in fact be distinguished with a finite-state model that is not a chain, but other examples require more sophisticated models. The best studied is probabilistic context-free grammar (PCFG), which operates over trees, categories of words, and individual lexical items, and has none of the restrictions of finite-state models. We find that PCFGs are state-of-the-art for parsing performance and are easier to learn from data than categorical context-free grammars. Other types of probabilistic models cover semantic and discourse structures. Every probabilistic model is a superset of a deterministic model (because the deterministic model could be seen as a probabilistic model where the probabilities are restricted to be 0 or 1), so any valid criticism of probabilistic models would have to be because they are too expressive, not because they are not expressive enough.

In *Syntactic Structures*, Chomsky introduces a now-famous example that is another criticism of finite-state probabilistic models:

> Neither (a) 'colorless green ideas sleep furiously' nor (b) 'furiously sleep ideas green colorless', nor any of their parts, has ever occurred in the past linguistic experience of an English speaker. But (a) is grammatical, while (b) is not.

Chomsky appears to be correct that neither sentence appeared in the published literature before 1955. I'm not sure what he meant by "any of their parts," but certainly every two-word part had occurred, for example:

- "It is neutral green, **colorless green**, like the glaucous water lying in a cellar." The Paris we remember, Elisabeth Finley Thomas (1942).
- "To specify those **green ideas** is hardly necessary, but you may observe Mr. [D. H.] Lawrence in the role of the satiated aesthete." The New Republic: Volume 29 p. 184, William White (1922).
- **"Ideas sleep** in books." Current Opinion: Volume 52, (1912).

But regardless of what is meant by "part," a statistically-trained finite-state model *can* in fact distinguish between these two sentences. Pereira (2001) showed that such a model, augmented with word categories and trained by expectation maximization on newspaper text, computes that (a) is 200,000 times more probable than (b). To prove that this was not the result of Chomsky's sentence itself sneaking into newspaper text, I repeated the experiment, using a much cruder model with Laplacian smoothing and no categories, trained over the Google Book corpus from 1800 to 1954, and found that (a) is about 10,000 times more probable. If we had a probabilistic model over trees as well as word sequences, we could perhaps do an even better job of computing degree of grammaticality.

Furthermore, the statistical models are capable of delivering the judgment that both sentences are *extremely* improbable, when compared to, say, "Effective green products sell well." Chomsky's theory, being categorical, cannot make this distinction; all it can distinguish is grammatical/ungrammatical.

Another part of Chomsky's objection is "we cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds." (Note that modern models are much larger than the 10^9 parameters that were contemplated in the 1960s.) But of course nobody is proposing that these parameters are learned one-by-one; the right way to do learning is to set large swaths of near-zero parameters simultaneously with a smoothing or regularization procedure, and update the high-probability parameters continuously as observations comes in. And noone is suggesting that Markov models by themselves are a serious model of human language performance. But I (and others) suggest that probabilistic, trained models are a better model of human language performance than are categorical, untrained models. And yes, it seems clear that an adult speaker of English does know billions of language facts (for example, that one says "big game" rather than "large game" when talking about an important football game). These facts must somehow be encoded in the brain.

It seems clear that probabilistic models are better for judging the likelihood of a sentence, or its degree of sensibility. But even if you are not interested in these factors and are only interested in the grammaticality of sentences, it still seems that probabilistic models do a better job at describing the linguistic facts. The *mathematical* theory of formal languages defines a language as a set of sentences. That is, every sentence is either grammatical or ungrammatical; there is no need for probability in this framework. But natural languages are not like that. A *scientific* theory of natural languages must account for the many phrases and sentences which leave a native speaker uncertain about their grammaticality (see Chris Manning's article and its discussion of the phrase "as least as"), and there are phrases which some speakers find perfectly grammatical, others perfectly ungrammatical, and still others will flip-flop from one occasion to the next. Finally, there are usages which are rare in a language, but cannot be dismissed if one is concerned with actual data. For example, the verb *quake* is listed as intransitive in dictionaries, meaning that (1) below is grammatical, and (2) is not, according to a categorical theory of grammar.

1. The earth quaked.
2. ? It quaked her bowels.

But (2) actually appears as a sentence of English. This poses a dilemma for the categorical theory. When (2) is observed we must either arbitrarily dismiss it as an error that is outside the bounds of our model (without any theoretical grounds for doing so), or we must change the theory to allow (2), which often results in the acceptance of a flood of sentences that we would prefer to remain ungrammatical. As Edward Sapir said in 1921, "All grammars leak." But in a probabilistic model there is no difficulty; we can say that *quake* has a high probability of being used intransitively, and a low probability of transitive use (and we can, if we care, further describe those uses through subcategorization).

Steve Abney points out that probabilistic models are better suited for modeling language change. He cites the example of a 15th century Englishman who goes to the pub every day and orders "Ale!" Under a categorical model, you could reasonably expect that one day he would be served eel, because the great vowel shift flipped a Boolean parameter in his mind a day before it flipped the parameter in the publican's. In a probabilistic framework, there will be multiple parameters, perhaps with continuous values, and it is easy to see how the shift can take place gradually over two centuries.

Thus it seems that grammaticality is not a categorical, deterministic judgment but rather an inherently probabilistic one. This becomes clear to anyone who spends time making *observations* of a corpus of actual sentences, but can remain unknown to those who think that the object of study is their own set of *intuitions* about grammaticality. Both observation and intuition have been used in the history of science, so neither is "novel," but it is observation, not intuition that is the dominant model for science.

Now let's consider what I think is Chomsky's main point of disagreement with statistical models: the tension between "accurate description" and "insight." This is an old distinction. Charles Darwin (biologist, 1809–1882) is best known for his insightful theories but he stressed the importance of accurate description, saying "False facts are highly injurious to the progress of science, for they often endure long; but false views, if supported by some evidence, do little harm, for every one takes a salutary pleasure in proving their falseness." More recently, Richard Feynman (physicist, 1918–1988) wrote "Physics can progress without the proofs, but we can't go on without the facts."

On the other side, Ernest Rutherford (physicist, 1871–1937) disdained mere description, saying "All science is either physics or stamp collecting." Chomsky stands with him: "You can also collect butterflies and make many observations. If you like butterflies, that's fine; but such work must not be confounded with research, which is concerned to discover explanatory principles."

Butterflies

Acknowledging both sides is Robert Millikan (physicist, 1868–1953) who said in his Nobel acceptance speech "Science walks forward on two feet, namely theory and experiment ... Sometimes it is one foot that is put forward first, sometimes the other, but continuous progress is only made by the use of both."

The two cultures

After all those distinguished scientists have weighed in, I think the most relevant contribution to the current discussion is the 2001 paper by Leo Breiman (statistician, 1928–2005), [Statistical Modeling: The Two Cultures]. In this paper Breiman, alluding to C.P. Snow, describes two cultures:

Leo Breiman

Leo Breiman

First the **data modeling culture** (to which, Breiman estimates, 98% of statisticians subscribe) holds that nature can be described as a black box that has a relatively simple underlying model which maps from input variables to output variables (with perhaps some random noise thrown in). It is the job of the statistician to wisely choose an underlying model that reflects the reality of nature, and then use statistical data to estimate the parameters of the model.

Second the **algorithmic modeling culture** (subscribed to by 2% of statisticians and many researchers in biology, artificial intelligence, and other fields that deal with complex phenomena), which holds that nature's black box cannot necessarily be described by a simple model. Complex algorithmic approaches (such as support vector machines or boosted decision trees or deep belief networks) are used to estimate the function that maps from input to output variables, but we have no expectation that the *form* of the function that emerges from this complex algorithm reflects the true underlying nature.

It seems that the algorithmic modeling culture is what Chomsky is objecting to most vigorously. It is not just that the models are statistical (or probabilistic), it is that they produce a form that, while accurately modeling reality, is not easily interpretable by humans, and makes no claim to correspond to the generative process used by nature. In other words, algorithmic modeling describes what *does* happen, but it doesn't answer the question of *why*.

Breiman's article explains his objections to the first culture, data modeling. Basically, the conclusions made by data modeling are about the model, not about nature. (Aside: I remember in 2000 hearing James Martin, the leader of the Viking missions to Mars, saying that his job as a spacecraft engineer was not to land on Mars, but to land on the model of Mars provided by the geologists.) The problem is, if the model does not emulate nature well, then the conclusions may be wrong. For example, linear regression is one of the most powerful tools in the statistician's toolbox. Therefore, many analyses start out with

"Assume the data are generated by a linear model..." and lack sufficient analysis of what happens if the data are not in fact generated that way. In addition, for complex problems there are usually many alternative good models, each with very similar measures of goodness of fit. How is the data modeler to choose between them? Something has to give. Breiman is inviting us to give up on the idea that we can uniquely model the true underlying *form* of nature's function from inputs to outputs. Instead he asks us to be satisfied with a function that accounts for the observed data well, and generalizes to new, previously unseen data well, but may be expressed in a complex mathematical form that may bear no relation to the "true" function's form (if such a true function even exists). Chomsky takes the opposite approach: he prefers to keep a simple, elegant model, and give up on the idea that the model will represent the data well. Instead, he declares that what he calls *performance* data—what people actually do—is off limits to linguistics; what really matters is *competence*—what he imagines that they should do.

Bill O'Reilly

Laplace

In January of 2011, television personality Bill O'Reilly weighed in on more than one culture war with his statement "*tide goes in, tide goes out. Never a miscommunication. You can't explain that*," which he proposed as an argument for the existence of God. O'Reilly was ridiculed by his detractors for not knowing that tides can be readily explained by a system of partial differential equations describing the gravitational interaction of sun, earth, and moon (a fact that was first worked out by Laplace in 1776 and has been considerably refined since; when asked by Napoleon why the creator did not enter into his calculations, Laplace said "I had no need of that hypothesis."). (O'Reilly also seems not to know about Deimos and Phobos (two of my favorite moons in the entire solar system, along with Europa, Io, and Titan), nor that Mars and Venus orbit the sun, nor that the reason Venus has no moons is because it is so close to the sun that there is scant room for a stable lunar orbit.) But O'Reilly realizes that it doesn't matter what his detractors think of his astronomical ignorance, because his supporters think he has gotten exactly to the key issue: *why?* He doesn't care *how* the tides work, tell him *why* they work. *Why* is the moon at the right distance to provide a gentle tide, and exert a stabilizing effect on earth's axis of rotation, thus protecting life here? *Why* does gravity work the way it does? *Why* does anything at all exist rather than not exist? O'Reilly is correct that these questions can only be addressed by mythmaking, religion or philosophy, not by science.

Chomsky has a philosophy based on the idea that we should focus on the deep *whys* and that mere explanations of reality don't matter. In this, Chomsky is in complete agreement with O'Reilly. (I recognize that the previous sentence would have an extremely low probability in a probabilistic model trained on a newspaper or TV corpus.) Chomsky believes a theory of language should be simple and understandable, like a linear regression model where we know the underlying process is a straight line, and all we have to do is estimate the slope and intercept.

For example, consider the notion of a pro-drop language from Chomsky's Lectures on Government and Binding (1981). In English we say, for example, "I'm hungry," expressing the pronoun "I". But in Spanish, one expresses the same thought with "Tengo hambre" (literally "have hunger"), dropping the pronoun "Yo". Chomsky's theory is that there is a "pro-drop parameter" which is "true" in Spanish and "false" in English, and that once we discover the small set of parameters that describe all languages, and the values of those parameters for each language, we will have achieved true understanding.

The problem is that reality is messier than this theory. Here are some dropped pronouns in English:

Dana Carvey

- "Not gonna do it. Wouldn't be prudent." (Dana Carvey, impersonating George H. W. Bush)
- "Thinks he can outsmart us, does he?" (Evelyn Waugh, The Loved One)
- "Likes to fight, does he?" (S.M. Stirling, The Sunrise Lands)
- "Thinks he's all that." (Kate Brian, Lucky T)
- "Go for a walk?" (countless dog owners)
- "Gotcha!" "Found it!" "Looks good to me!" (common expressions)

Linguists can argue over the interpretation of these facts for hours on end, but the diversity of language seems to be much more complex than a single Boolean value for a pro-drop parameter. We shouldn't accept a theoretical framework that places a priority on making the model simple over making it accurately reflect reality.

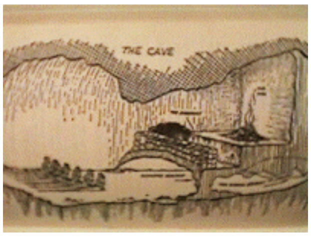
From the beginning, Chomsky has focused on the *generative* side of language. From this side, it is reasonable to tell a non-probabilistic story: I *know* definitively the idea I want to express—I'm starting from a single semantic form—thus all I have to do is choose the words to say it; why can't that be a deterministic, categorical process? If Chomsky had focused on the other side, *interpretation*, as Claude Shannon did, he may have changed his tune. In interpretation (such as speech recognition) the listener receives a noisy, ambiguous signal and needs to decide which of many possible intended messages is most likely. Thus, it is obvious that this is inherently a probabilistic problem, as was recognized early on by all researchers in speech recognition, and by scientists in other fields that do interpretation: the astronomer Laplace said in 1819 "Probability theory is nothing more than common sense reduced to calculation," and the physicist James Maxwell said in 1850 "The true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind."

Finally, one more reason why Chomsky dislikes statistical models is that they tend to make linguistics an empirical science (a science about how people actually use language) rather than a mathematical science (an investigation of the mathematical properties of models of formal language). Chomsky prefers the later, as evidenced by his statement in *Aspects of the Theory of Syntax* (1965):

> Linguistic theory is mentalistic, since it is concerned with discovering a mental reality underlying actual behavior. Observed use of language ... may provide evidence ... but surely cannot constitute the subject-matter of linguistics, if this is to be a serious discipline.

I can't imagine Laplace saying that observations of the planets cannot constitute the subject-matter of orbital mechanics, or Maxwell saying that observations of electrical charge cannot constitute the subject-matter of electromagnetism. It is true that physics considers idealizations that are abstractions from the messy real world. For example, a class of mechanics problems ignores friction. But that doesn't mean that friction is not considered part of the subject-matter of physics.

So how could Chomsky say that observations of language cannot be the subject-matter of linguistics? It seems to come from his viewpoint as a Platonist and a Rationalist and perhaps a bit of a Mystic. As in Plato's allegory of the cave, Chomsky thinks we should focus on the ideal, abstract forms that underlie language, not on the superficial manifestations of language that happen to be perceivable in the real world. That is why he is not interested in language performance. But Chomsky, like Plato, has to answer where these ideal forms come from. Chomsky (1991) shows that he is happy with a Mystical answer, although he shifts vocabulary from "soul" to "biological endowment."



Plato's cave

> Plato's answer was that the knowledge is 'remembered' from an earlier existence. The answer calls for a mechanism: perhaps the immortal soul ... rephrasing Plato's answer in terms more congenial to us today, we will say that the basic properties of cognitive systems are innate to the mind, part of human biological endowment.

It was reasonable for Plato to think that the ideal of, say, a horse, was more important than any individual horse we can perceive in the world. In 400BC, species were thought to be eternal and unchanging. We now know that is not true; that the horses on another cave wall—in Lascaux—are now extinct, and that current horses continue to evolve slowly over time. Thus there is no such thing as a single ideal eternal "horse" form.


Lascaux Horse

We also now know that language is like that as well: languages are complex, random, contingent biological processes that are subject to the whims of evolution and cultural change. What constitutes a language is not an eternal ideal form, represented by the settings of a small number of parameters, but rather is the contingent outcome of complex processes. Since they are contingent, it seems they can only be analyzed with probabilistic models. Since people have to continually understand the uncertain. ambiguous, noisy speech of others, it seems they must be using something like probabilistic reasoning. Chomsky for some reason wants to avoid this, and therefore he must declare the actual facts of language use out of bounds and declare that true linguistics only exists in the mathematical realm, where he can impose the formalism he wants. Then, to get language from this abstract, eternal, mathematical realm into the heads of people, he must fabricate a mystical facility that is exactly tuned to the eternal realm. This may be very interesting from a mathematical point of view, but it misses the point about what language is, and how it works.

Thanks

Thanks to Ann Farmer, Fernando Pereira, Dan Jurafsky, Hal Varian, and others for comments and suggestions on this essay.

Annotated Bibliography

1. Abney, Steve (1996) Statistical Methods and Linguistics, in Klavans and Resnik (eds.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press.

   An excellent overall introduction to the statistical approach to language processing, and covers some ground that is not addressed often, such as language change and individual differences.

2. Breiman, Leo (2001) Statistical Modeling: The Two Cultures, *Statistical Science*, Vol. 16, No. 3, 199-231.

   Breiman does a great job of describing the two approaches, explaining the benefits of his approach, and defending his points in the vary interesting commentary with eminent statisticians: Cox, Efron, Hoadley, and Parzen.

3. Chomsky, Noam (1956) Three Models for the Description of Language, *IRE Transactions on Information theory* (2), pp. 113-124.

   Compares finite state, phrase structure, and transformational grammars. Introduces "colorless green ideas sleep furiously."

4. Chomsky, Noam (1967) Syntactic Structures, Mouton.

   A book-length exposition of Chomsky's theory that was the leading exposition of linguistics for a decade. Claims that probabilistic models give no insight into syntax.

5. Chomsky, Noam (1969) Some Empirical Assumptions in Modern Philosophy of Language, in *Philosophy, Science and Method: Essays in Honor or Ernest Nagel*, St. Martin's Press.

   Claims that the notion "probability of a sentence" is an entirely useless notion.

6. Chomsky, Noam (1981) Lectures on government and binding, de Gruyer.

   A revision of Chomsky's theory; this version introduces Universal Grammar. We cite it for the coverage of parameters such as pro-drop.

7. Chomsky, Noam (1991) Linguistics and adjacent fields: a personal view, in Kasher (ed.), *A Chomskyan Turn*, Oxford.

I found the Plato quotes in this article, published by the Communist Party of Great Britain, and apparently published by someone with no linguistics training whatsoever, but with a political agenda.

8. Gold, E. M. (1967) Language Identification in the Limit, *Information and Control*, Vol. 10, No. 5, pp. 447-474.

   Gold proved a result in formal language theory that we can state (with some artistic license) as this: imagine a game between two players, guesser and chooser. Chooser says to guesser, "Here is an infinite number of languages. I'm going to choose one of them, and start reading sentences to you that come from that language. On your N-th birthday there will be a True-False quiz where I give you 100 sentences you haven't heard yet, and you have to say whether they come from the language or not." There are some limits on what the infinite set looks like and on how the chooser can pick sentences (he can be deliberately tricky, but he can't just repeat the same sentence over and over, for example). Gold's result is that if the infinite set of languages are all generated by context-free grammars then there is no strategy for guesser that guarantees she gets 100% correct every time, no matter what N you choose for the birthday. This result was taken by Chomsky and others to mean that it is impossible for children to learn human languages without having an innate "language organ." As Johnson (2004) and others show, this was an invalid conclusion; the task of getting 100% on the quiz (which Gold called language identification) really has nothing in common with the task of language acquisition performed by children, so Gold's Theorem has no relevance.

9. Horning, J. J. (1969) A study of grammatical inference, Ph.D. thesis, Stanford Univ.

   Where Gold found a negative result—that context-free languages were not identifiable from examples, Horning found a positive result—that probabilistic context-free languages are identifiable (to within an arbitrarily small level of error). Nobody doubts that humans have unique innate capabilities for understanding language (although it is unknown to what extent these capabilities are specific to language and to what extent they are general cognitive abilities related to sequencing and forming abstractions). But Horning proved in 1969 that Gold cannot be used as a convincing argument for an innate language organ that specifies all of language except for the setting of a few parameters.

10. Johnson, Kent (2004) Gold's Theorem and cognitive science, *Philosophy of Science*, Vol. 71, pp. 571-592.

    The best article I've seen on what Gold's Theorem actually says and what has been claimed about it (correctly and incorrectly). Concludes that Gold has something to say about formal languages, but nothing about child language acquisition.

11. Lappin, Shalom and Shieber, Stuart M. (2007) Machine learning theory and practice as a source of insight into universal grammar., *Journal of Linguistics*, Vol. 43, No. 2, pp. 393-427.

    An excellent article discussing the poverty of the stimulus, the fact that all models have bias, the difference between supervised and unsupervised learning, and modern (PAC or VC) learning theory. It provides alternatives to the model of Universal Grammar consisting of a fixed set of binary parameters.

12. Manning, Christopher (2002) Probabilistic Syntax, in Bod, Hay, and Jannedy (eds.), *Probabilistic Linguistics*, MIT Press.

    A compelling introduction to probabilistic syntax, and how it is a better model for linguistic facts than categorical syntax. Covers "the joys and perils of corpus linguistics."

13. Norvig, Peter (2007) How to Write a Spelling Corrector, unpublished web page.

Shows working code to implement a probabilistic, statistical spelling correction algorithm.

14. Norvig, Peter (2009) Natural Language Corpus Data, in Seagran and Hammerbacher (eds.), *Beautiful Data*, O'Reilly.

Expands on the essay above; shows how to implement three tasks: text segmentation, cryptographic decoding, and spelling correction (in a slightly more complete form than the previous essay).

15. Pereira, Fernando (2002) Formal grammar and information theory: together again?, in Nevin and Johnson (eds.), *The Legacy of Zellig Harris*, Benjamins.

When I set out to write the page you are reading now, I was concentrating on the events that took place in Cambridge, Mass., 4800 km from home. After doing some research I was surprised to learn that the authors of two of the three best articles on this subject sit within a total of 10 meters from my desk: Fernando Pereira and Chris Manning. (The third, Steve Abney, sits 3700 km away.) But perhaps I shouldn't have been surprised. I remember giving a talk at ACL on the corpus-based language models used at Google, and having Fernando, then a professor at U. Penn., comment "I feel like I'm a particle physicist and you've got the only super-collider." A few years later he moved to Google. Fernando is also famous for his quote "The older I get, the further down the Chomsky Hierarchy I go." His article here covers some of the same ground as mine, but he goes farther in explaining the range of probabilistic models available and how they are useful.

16. Plato (c. 380BC) The Republic.

Cited here for the allegory of the cave.

17. Shannon, C.E. (1948) A Mathematical Theory of Communication, *The Bell System Technical Journal*, Vol. 27, pp. 379-423.

An enormously influential article that started the field of information theory and introduced the term "bit" and the noisy channel model, demonstrated successive n-gram approximations of English, described Markov models of language, defined entropy with respect to these models, and enabled the growth of the telecommunications industry.

Peter Norvig
