By this species of argument, stochastic models are practically always a stop-gap approximation. Take stochastic queue theory, for example, by which one can give a probabilistic model of how many trucks will be arriving at given depots in a transportation system. One could argue that if we could just model everything about the state of the trucks and the conditions of the roads, the location of every nail that might cause a flat and every drunk driver that might cause an accident, then we could in principle predict deterministically how many trucks will be arriving at any depot at any time, and there is no need of stochastic queue theory. Stochastic queue theory is only an approximation in lieue of information that it is impractical to collect.

But this argument is flawed. If we have a complex deterministic system, and if we have access to the initial conditions in complete detail, so that we can compute the state of the system unerringly at every point in time, a simpler stochastic description may still be more insightful. To use a dirty word, some properties of the system are genuinely *emergent*, and a stochastic account is not just an approximation, it provides more insight than identifying every deterministic factor. Or to use a different dirty word, it is a *reductionist* error to reject a successful stochastic account and insist that only a more complex, lower-level, deterministic model advances scientific understanding.

4.2   Chomsky v. Shannon

In one's introductory linguistics course, one learns that Chomsky disabused the field once and for all of the notion that there was anything of interest to statistical models of language. But one usually comes away a little fuzzy on the question of what, precisely, he proved.

The arguments of Chomsky's that I know are from "Three Models for the Description of Language" [5] and *Syntactic Structures* [6] (essentially the same argument repeated in both places), and from the *Handbook of Mathematical Psychology*, chapter 13 [17]. I think the first argument in *Syntactic Structures* is the best known. It goes like this.

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

This argument only goes through if we assume that if the frequency of a sentence or 'part' is zero in a training sample, its probability is zero. But in fact, there is quite a literature on how to estimate the probabilities of events that do not occur in the sample, and in particular how to distinguish real zeros from zeros that just reflect something that is missing by chance.

Chomsky also gives a more general argument:

> If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list; there appears to be no particular relation between order of approximation and grammaticalness.

Because for any n, there are sentences with grammatical dependencies spanning more than n words, so that no nth-order statistical approximation can sort out the grammatical from the ungrammatical examples. In a word, you cannot define grammaticality in terms of probability.

It is clear from context that 'statistical approximation to English' is a reference to nth-order Markov models, as discussed by Shannon. Chomsky is saying that there is no way to choose n and ϵ such that

for all sentences s, grammatical$(s) \leftrightarrow P_n(s) > \epsilon$

where $P_n(s)$ is the probability of s according to the 'best' nth-order approximation to English.

But Shannon himself was careful to call attention to precisely this point: that for any n, there will be some dependencies affecting the well-formedness of a sentence that an nth-order model does not capture. The point of Shannon's approximations is that, as n increases, the total mass of ungrammatical sentences that are erroneously assigned nonzero probability decreases. That is, we *can* in fact define grammaticality in terms of probability, as follows:

grammatical$(s) \leftrightarrow \lim_{n \to \infty} P_n(s) > 0$

A third variant of the argument appears in the *Handbook*. There Chomsky states that parameter estimation is impractical for an nth-order Markov model where n is large enough "to give a reasonable fit to ordinary usage". He emphasizes that the problem is not just an inconvenience for statisticians, but renders the model untenable as a model of human language acquisition: "we cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds."

This argument is also only partially valid. If it takes at least a second to estimate each parameter, and parameters are estimated sequentially, the argument is correct. But if parameters are estimated in parallel, say, by a high-dimensional iterative or gradient-pursuit method, all bets are off. Nonetheless, I think even the most hardcore statistical types are willing to admit that Markov models represent a brute force approach, and are not an adequate basis for psychological models of language processing.

However, the inadequacy of Markov models is not that they are statistical, but that they are statistical versions of finite-state automata! Each of Chomsky's arguments turns on the fact that Markov models are finite-state, not on the fact that they are stochastic. None of his criticisms are applicable

to stochastic models generally. More sophisticated stochastic models do exist: stochastic context-free grammars are well understood, and stochastic versions of Tree-Adjoining Grammar [18], GB [8], and HPSG [3] have been proposed.

In fact, probabilities make Markov models more adequate than their non-probabilistic counterparts, not less adequate. Markov models are surprisingly effective, given their finite-state substrate. For example, they are the workhorse of speech recognition technology. Stochastic grammars can also be easier to learn than their non-stochastic counterparts. For example, though Gold [9] showed that the class of context-free grammars is not learnable, Horning [13] showed that the class of stochastic context-free grammars *is* learnable.

In short, Chomsky's arguments do not bear at all on the probabilistic nature of Markov models, only on the fact that they are finite-state. His arguments are not by any stretch of the imagination a sweeping condemnation of statistical methods.

5   Conclusion

In closing, let me repeat the main line of argument as concisely as I can. Statistical methods—by which I mean primarily weighted grammars and distributional induction methods—are clearly relevant to language acquisition, language change, language variation, language generation, and language comprehension. Understanding language in this broad sense is the ultimate goal of linguistics.

The issues to which weighted grammars apply, particularly as concerns perception of grammaticality and ambiguity, one may be tempted to dismiss as performance issues. However, the set of issues labelled "performance" are not essentially computational, as one is often led to believe. Rather, "competence" represents a provisional narrowing and simplification of data in order to understand the algebraic properties of language. "Performance" is a misleading term for "everything else". Algebraic methods are inadequate for understanding many important properties of human language, such as the measure of goodness that permits one to identify the correct parse out of a large candidate set in the face of considerable noise.

Many other properties of language, as well, that are mysterious given unweighted grammars, properties such as the gradualness of rule learning, the gradualness of language change, dialect continua, and statistical universals, make a great deal more sense if we assume weighted or stochastic grammars. There is a huge body of mathematical techniques that computational linguists have begun to tap, yielding tremendous progress on previously intransigent problems. The focus in computational linguistics has admittedly been on technology. But the same techniques promise progress at long last on questions about the nature of language that have been mysterious for so long. The time is ripe to apply them.