Towards a new empiricism

John Goldsmith
The University of Chicago

1 Introduction

My purpose in this paper is to sketch some reasons for a linguist to take seriously an empiricist perspective of science.[1] This will involve reviewing the history of the term empiricism, its uses and abuses, and then reconsidering the utility of one of the crucial elements of modern empiricist epistemology: the notion of probability. I will suggest, on the one hand, that probability can be understood as the *quantitative theory of evidence*, and, on the other hand, that probability theory offers us a more comprehensive and concrete way to understand the fundamental problem of induction, which is to say, how one passes from knowledge of a finite number of particulars to a generalization, which, typically, allows us to infer an infinite number of conclusions, almost all of which have not yet been tested.[2] I will also plead for a non-cognitive conception of linguistics, and argue that the proper (or at least *a* proper) level of abstraction for linguistic theory is one that is more abstract than one that sees itself as a theory of the brain. But I will not plead for an *inductive* view of science: science remains the field *par excellence* in which progress can only be made by the creative efforts of those who engage in it, and those creative efforts can only be evaluated in the light of the theoretical insights they provide and the tightening of the empirical fit with the data.

---

[1] This is a draft (number 1.6, on my count), and comments are most welcome; please address them to goldsmith@uchicago.edu. An earlier version has appeared in *Recherches linguistiques à Vincennes* 36 (2007), edited by Joaquim Brandão de Carvalho, and if all goes well, a much revised version will be incorporated into a book that will be co-written with Alex Clark, Nick Chater, and Amy Perfors, to appear in late 2009.

[2] One of the consequences of this perspective is the perhaps surprising principle that the value of a theoretical innovation is neither more nor less than the amount of information it would take to (merely) stipulate its intended consequences.

2   Empiricism

2.1   Some historical observations

The term *empiricism* has covered a wide variety of views regarding the nature of mind and the ways in which we can best learn about the natural world around us. In its early days, it was associated with the views of John Locke and David Hume, and during the 17th and 18th centuries, empiricists differed in some important ways with their rationalist colleagues on the Continent with regard to the origin of human knowledge. Empiricists and rationalists were largely on the same side of the important intellectual battles of the age, which involved the importance of scientific investigation as the most important way to understand the world. Both empiricists and rationalists were in agreement that science was a better authority than either the Aristotelian scholastics or the established churches of their time were.

Empiricists and rationalists differed with regard to whether the careful study of the information that is presented to the human mind is sufficient to account for all of human knowledge. Empiricists took the strongest possible stand: all knowledge comes through the senses. Rationalists disagreed, and argued either that the empiricist position was *necessarily* wrong, because the learning process cannot start from nothing—or that the empiricist position was *simply* wrong, because we do know things that the senses alone could not convince us of.

Empiricists' central example of what one knows is one's own report of what one sees, tastes, or hears, while rationalists' central example of what one knows is a mathematical statement of which one knows well the proof (my own favorite example is that a group of prime order has no proper subgroups, but if that does not appeal to the reader, then one could take a different example, like the statement that the only even prime number is 2). For their part, rationalists weren't entirely certain that we really *know* what we think we perceive; we could be mistaken—while empiricists believed that rationalists were ignoring and overlooking some uncertainties with regard to mathematics that simply couldn't be held against sensory reports.

Bear in mind the following: in the 17th and 18th century, there was no distinction to be drawn between philosophy and psychology, and this lack of distinction can be treacherous for us, as we look back on the issues of those days, because we, in the 21st century, make a distinction that seems to us rather evident. On the one hand, we can study what it is for *us* to know something, a question that today calls upon the disciplines of philosophy, history, and sociology, but on the other hand, we can also study what it is for *humans* to know something, a question that we typically hand over

to psychologists, neuroscientists, and their colleagues. Never mind, at least for now, that there is a great overlap between being us and being human; I want merely to draw the distinction in how we ask about knowledge—its foundations, its origins, its trustworthiness. (There is a tradition that tries, as best it can, to erase the distinction that I am trying to clarify, a tradition that is often referred to as cognitive naturalism, or the like, often associated with Willard Quine's later views.) I will refer to the question of what it is for us to know something as the *philosopher's* question of knowledge (or the *epistemological* question), and to the question of what it means for a human being to know something as the *psychologist's* question of knowledge.[3]

The philosopher's question of knowledge is posed starkly by Descartes: how can I establish my knowledge of anything if I start by doubting everything, including what my senses tell me? Descartes' own answer is based on a search from inside his mind, and the early modern tradition of the 17th and 18th century followed that path as well. By the 19th century, answers to this question began to come from without: from studies of the history of ideas, the history of the material world, and the history of the political world.

The psychologist's question of knowledge only became clear when it became possible to ask questions of interest about the knowledge of other people by methods that were not different in kind from methods used to study other things in the natural world, and this moment was the beginning of the discipline of psychology. Its roots, as we have said, lay in philosophical work, but the methods that created a modern psychology came into their own in the middle of the 19th century. This disciplinary evolution was also the result of the impact of Darwin's theory of evolution, and the conclusion that human beings were, from a biological point of view, part of the same natural world as apes, rats, and pigeons.

We move forward to the 20th century. Virtually all of the issues that separated rationalists and empircists have been so morphed by the passage of time and the advances of science that the terms are largely archaisms. But not entirely: the term *empiricism* came back into vogue early in the 20th

---

[3]If you are wondering how we can make such a distinction—since we really do know, individually, that we are human beings—we will certainly return to this question, but for the moment, simply consider what conclusion you would draw if (to take a not unreasonable example) it was shown to you that psychological methods established that you had no awareness of yourself, or were not able to draw rational conclusions—in short, some conclusion which you had personal, but not scientific, reasons to dismiss. You would, in such a case, compartmentalize the scientific conclusion from the principles you employ during your normal life. I will return to this shortly when I bring up the charges that have been leveled against the modern empiricists for seeing the mind as too *malleable.*

century, largely due to its use in Berlin by a group that called itself logical empiricists, and the term's association with the influential Vienna Circle, which called its views logical positivism.[4] The term *rationalism* came back into vogue later in the century largely due to its use by Chomsky and other cognitive scientists, beginning in the 1960s. I will refer to these as *modern* empiricism and rationalism, unless context makes it so clear that we may omit the term *modern.*

Modern empiricism's principal target was the nature of scientific knowledge and knowledge *tout court*—an epistemological question. Empiricism held close to the positivist movement in seeing scientific knowledge as the highest model of knowledge. In some respects, empiricism was a wipe-the-slate-clean philosophy, much like Descartes's general doubting, or Socrates's "I only know that I know nothing." In its most extravagant form, psychological behaviorism, it made strong claims of dubious merit (though we should bear in mind that behaviorism predates modern empiricism by a decade, or two; its official date of birth is 1913). In its more constructive form, it offered a philosophical justification for a freedom of spirit and thought which allowed young scientists to overthrow the heavy weight of a moribund tradition, and to replace it with such new frameworks as the theories of relativity and of quantum mechanics.

Modern rationalism is not primarily a theory of epistemology; it is an approach to understanding how human beings think and know.[5] It is a child

---

[4]The development of the formal tools used to characterize syntax—and thus the modern generative framework—grew out of the work of the Vienna Circle, and others directly influenced by them, such was Willard Quine. See [12] for extended discussion. Logical empiricism, as a movement, was based on the desire to clarify precisely what the sensorial basis of scientific knowledge was. It invested heavily in the analysis of language because sensorial evidence is notoriously private and ineffable, and it is not at all obvious how a philosophy of science that views all knowledge as based on experience can achieve a foundation for scientific knowledge that is not ineffable. A major part of the effort along these lines consisted, first, in trying to establish a way of speaking about scientific matters that was as close to simple report of sensory experience as possible, and second, in clarifying what aspects of a theory's pronouncements are matters of definition or logic. The classic postulates of empiricism (against which Quine argued in [19]) were (1) that all true statements can be divided into those that are true by virtue of the meanings of the terms–called analytic–and those that are true by virtue of observations–called synthetic (the clear analytic/synthetic distinction), and (2) that all theoretical statements can be broken down into statements ultimately about observations (so-called reductionism)

[5]It is true, however, that in a range of publications, Chomsky has tried to make the case for an idiosyncratic view of science; it is one in which data is taken seriously only to the extent that the scientist feels it ought to be. A typical characterization is this, from [7], p. 19:

As regards data, the standpoint of the natural sciences is instrumental. Data

of the mathematical breakthroughs of the 1930s, 1940s, and 1950s involving a new conception of logical inference, of mathematical proof, and ultimately of how something resembling thought could be embodied in what we today call a computer.

The empiricist perspective has often been seen to be allied with a fundamentally optimistic view of the human condition, which seems to me quite justified. Empiricism is, at the very least, optimistic about the possibility of better understanding the mind by better understanding its capacity to *learn*, where the learning in question is learning *from experience*. This seems to me to be a good thing to be optimistic about. If we humans fail in some attempt, we can at least face tomorrow with some hope, as long as we believe that we are capable of learning something from our failure, something that will help us to overcome our problems tomorrow in a new way, one which may be better than the one that has just failed us.

Curiously, the modern rationalist characterizes this optimism in a quite different light, one which views the mind as "malleable"—as if that were a disreputable thing to be. Jan Koster [15], for example, remarks:[6]

> Ideas of an initially empty and malleable human mind have always been considered progressive and were all too obviously ideologically motivated, as was pointed out by Harry Bracken in several books and articles (for instance, Bracken 1984 [2]).

To say a mind is "malleable" is to suggest just a bit more than that the mind can undergo change, since any mind that learns must by that simple

---

are important only insofar as they provide *evidence*, that is, evidence *for*, a relational notion. The goal is to find evidence that provides insight into a hidden reality. Adopting this approach, we abandon the hopeless search for order in the world of diret experience, and regard what can be observed as a means to gain access to the inner mechanisms of mind.

Such a view has nothing to do with rationalism, and is not at all a healthy perspective from the point of view of the research scientist. I will return to this below.

[6]It might be that Chomsky's remarks in *Language and Responsibility* [6] are the source of these observations:

> This set of [empiricist] beliefs corresponds very well to the demands of the technocratic intelligentsia: it offers them a very important social role. And in order to justify such practices, it is very useful to believe that human beings are empty organisms, malleable, controllable, easy to govern, and so on, with no essential need to struggle to find their own way and to determine their own fate. For that empiricism is quite suitable. So from this point of view, it is perhaps no surprise that denial of any "essential human nature" has been so prominent in much of left-wing doctrine.

fact change. But to be malleable is to change along lines designed by someone else, for ends that are likely hidden and very possible not in the best interest of the person involved. Such rhetoric! If we want to understand how minds learn, we must endow them with the ability to change with experience. The ability to learn is the gift of being malleable in the right way.[7]

To the extent that a modern-day rationalist posits little room for minds to learn from experience, that rationalist must make a decision as to whether this pessimism about how human beings learn (or fail to learn) will be extended to a general theory of epistemology, which is to say, whether this pessimism will come back and force itself on that rationalist as a condition on possible scientific theories. Is the scientist (who cannot use much more than his mind, after all) similarly limited with respect to what hypotheses he can entertain? Is the scientist capable of learning from observations and reflections? To the extent that the modern-day rationalist puts little stock in the ability of the humans he studies to learn from experience, but great stock in his own ability to craft a fine theory from his experience—to that extent he owes us some serious explanations of his apparent inconsistency. To be sure, he *can* deny science the ability to learn from experience too, but at that point, some of us stop listening and go back to work.

2.2   The problem of induction

There is one concern that the philosophical and the psychological approaches share: this is what is known as the problem of induction, which is the question of how one can legitimately pass from knowledge of a finite number of particulars to knowledge of a generalization, especially generalizations which, like most generalizations, have an infinite number of consequences which follow from them. From the philosopher's point of view, any such inference is tricky business, since it is reasonably clear that there is a serious danger of making a mistake and drawing the *wrong* conclusion from the observations. From the psychologist's point of view, understanding how people (or for that matter, rats) draw inferences is a difficult task, but one which holds the promise of telling us something about otherwise scientifically inaccessible reaches of the human psyche.[8]

---

[7]If we don't go down that road, we may end up with a philosophy in which every concept must be innate, a philosophy which is its own refutation.

[8]In passing, I would like to note that *this* was the the point that excited early behaviorists, for whom Darwin's view of man as part of the biological world was still fresh and vital; the behaviorists were the first psychologists to make it part of their manifesto that what we learn about animal learning is certain to tell us about human learning, based on Darwin's insights.

From the linguist's point of view, the problem of induction is an abstract way to formulate the most central question of synchronic linguistic analysis: how do we justify the formulation of grammatical statements, valid for a potentially infinite class of representations, on the basis of a finite set of data? The current Chomskian ("principles and parameters") approach is based on the hope that only one element in a restricted class of grammars, those permitted by UG, are consistent with the observed data. The empiricist account is that a careful study of the formal and statistical properties of observable data will lead to a "rating" of grammars which generate the data calculated from a probabilistic model, and that the most probable model is the best one. This paper is an attempt to spell that out in more detail.

Oddly enough, as we will see, I believe that what I will describe here is very much in the tradition of classical generative grammar—by which I mean, the research program described in detail in *The Logical Structure of Linguistic Theory* [5], but this line of thinking does make one quite skeptical about the principles and parameters approach to grammar. Be prepared to jettison some beliefs.

2.3 Empiricism and linguistics

Linguists in the pre-generative period in American linguistics would happily have referred to themselves as empiricists. Foremost in their understanding of the term was the sense that empiricists were highly skeptical of what they viewed as metaphysical claims, and they shared the view that a rigorous method needed to be invoked in order to make scientific claims.

For that reason, linguists expended considerable time, energy, and attention discussing and developing notions of linguistic method. To many generative linguists today, this earlier emphasis on method, along with an apparent lack of concern for something else called "theory," makes empiricist views seem more mysterious than they really are. Empiricist methods have at their core two rough-and-ready principles: first, that the data are what they are, not what the linguist wants them to be, and second, that care must be taken to justify the positing of abstract entities in one's theoretical analysis—or to put the matter another way, while it is fine to be proud to have discovered an unseen object, the burden of proof remains heavy on the scientist who claims to have found one. A direct consequence of this is that alternative analyses in which abstract elements are not posited have to be thoroughly explored to be sure that none of them is as capable of accounting for the evidence.

The abstract element that I would like to skeptically rethink in this paper is Universal Grammar—UG, for short. UG is a central concept in much

current thinking about linguistics; it is sometimes conceptualized as the initial state of the human language acquisition device; it is, in any event, the conceptual substance necessary to bridge the gap between the linguistic data presented to the child-learner and the grammar that he ends up with as a fully competent adult native speaker. The question is not so much whether UG exists, as it is to what extent UG should be thought of as conceptually rich. The empiricist assumption regarding human knowledge in the golden days of empiricism in the 17th and 18th century was that the mind starts off like a tabula rasa: a blank white-board, so to speak, on which drawings of any kind, and formulas in any language, could be written; UG, on this account, is relatively impoverished. The opposing picture today is one in which the mind starts off more like the control panel of a jet airliner, with a panoply of gauges and dials which need to be set, but whose settings only gain sense and meaning by virtue of the circuitry that lies behind the dashboard.

The main question I will attack in this paper is whether the role played by Universal Grammar can be assigned to a completely abstract and, we might say, platonic object, one that is based on algorithmic complexity, and unrelated to any particular biological characteristics of human beings. I will argue that such a project is feasible, given our present knowledge. In fact, the perspective I am describing here has a stronger claim to being called "universal grammar" than UG does, in the sense that the empiricist position described here would be valid in any spot in the known universe, and is not a theory of the human genetic endowment. [9]

---

[9]Regardless of one's views about linguistics and about language, it must be acknowledged that skepticism with regarding to positing abstract entities is a deep and abiding theme in modern thought, and lies at the heart of the Scientific Revolution begun by Galileo and Descartes. They led a battle, in particular, against the Scholastic tradition's view of the world in which objects were possessed of a range of qualities which seem difficult for the modern reader to take seriously—laudanum makes us sleepy because it has a dormitive property, for example. It may make us sleepy, but today we do not take seriously the positing of such a quality: we consider that mere verbal play. (If you think we are immune to that fallacy today, consider how explanatory an OT constraint like *Coda really is, and go from there.) Galileo and Descartes pushed for a much simpler world view, one in which what really exists is only quantity and spatial measure—a simplicity and economy of thought that could easily be linked to a kind of mathematics which was being developed at the time. All physical effects were local, and consisted of one object coming into contact with another. No longer would philosophers talk about objects as having their "natural place" at the center of the Earth, as the Scholastics taught; objects simply did not have natural places anymore, nor hidden qualities that were supposed to explain manifest behavior. As Newton wrote in *Optics*:

> To explain each property of things by endowing it with a specific occult quality by which the effects which we see are generated and produced is not to explain anything at all.(cited in Duhem 1903 [9] , p. 34, my translation.)

If rationalists were less skeptical than empiricists about the innate properties of the mind, rationalists were fervent skeptics of metaphysical claims about the natural world—like their empiricist colleagues, only more so. Indeed, the rationalist world (composed in large part, in 17th century Europe, of Cartesians and atomists) looked quite askance at Newton's theory of universal gravity, because it seemed to them to be a big step backwards, towards the Scholastic ways of thinking and the positing of occult "virtues" (properties, qualities). Huygens wrote (11 July 1687) in a letter,

> I would like to read Mr. Newton's book. It's fine with me if he's not a Cartesian, just so long as he doesn't make us swallow suppositions like forces of attraction. (op. cit.p. 33.)

Leibniz, the leading rationalist of the day, wrote to Huygens in an exasperated tone, seeing Newton's proposal as a step backwards to the methods of the Scholastics, with their "virtues":

> I do not understand how he conceives of weight or attraction. It seems that from his point of view, there is just an incorporeal and inexplicable virtue.(op.cit., p. 33.)

Huygens, three years later, wrote much the same again to Leibniz:

> As for what Mr. Newton takes the cause of flow to be, I am not at all happy, and I feel that way about all his other theories which he builds on his principle of attraction, which seems to me to be absurd.(op. cit., p.33.)

Newton's view was different. He shared with his rationalist critics a concern with positing a new sort of object in the world — what he called force — especially in light of the fact that so much of the Scholastic mumbo-jumbo had been eradicated from the philosophical vocabulary. He would much rather have had a theory of gravity which acted locally, like everything did in Descartes' view of the world. But Newton knew that the numbers didn't lie, and the comment I cited above by Newton is followed by his description of how he viewed his work: "But to draw from the phenomena two or three general principles of motion, and then to explain all the particularities of bodies by means of these clear principles—this is truly, in philosophy, to make great progress, even if the causes of these principles are not yet discovered; this is why I do not hesitate to propose the laws of motion, all the while leaving aside the search for its causes."(op. cit., pp. 34-5.) Other empiricists attacked back; Roger Cotes, in his introduction to the second edition of Newton's *Principe*, railed against the extent to which rationalists would hypothesize odd shapes and hypothetical fluids in order to account for observed behavior in a totally local fashion: "the story that they spin is elegant and pretty, but it is nothing more than a fable."(op. cit., p. 35.)

And Cotes was right. Still, one has to be careful in drawing ultimate conclusions from this story. Newton was also right, of course, both in what he proposed and in his hesitation in accepting a non-local theory of gravitation, but it would take the development of the notion of a field in physics, in the 19th century, and then the notion of space-time in the early 20th century, to find a way to rethink Newton's theory of gravity in a purely geometric way. And so, in a certain sense, the rationalists were right, too, about being skeptical; but if their skepticism had won the day, it is unlikely that physics would ever have gotten to the general theory of relativity, the only theory that could satisfy them.

3   A new empiricism

I will turn now to the central positive point of this paper, the heart of what I will call the new empiricism. It begins with the observation that the central problem of empirical knowledge, and hence of science, is the problem of induction: what justifies the passage of belief from a finite number of observations to a generalization which has an infinite number of testable consequences?

The answer is that a justifiable generalization is a probabilistic one: it assigns a non-negative probability to every predicted outcomes, in such a fashion that the probabilities sum to 1.0—neither more, nor less. In most cases of interest, the number of predicted outcomes is infinite, but the same condition holds: the sum of the probabilities of each outcome must be 1.0. For this condition to hold over an infinite set, it must be the case, first of all, that the probabilities get indefinitely small, and it must also be true that while we cannot test every outcome (since there are an infinite number of them), we can always find a finite number of outcomes whose total probability gets arbitrarily close to the probability of the whole set. Thus a probability measure assigned to an infinite set makes it *almost* as manageable as a finite set, while still remaining resolutely infinite.That is the heart of the matter.

Perhaps I should make it clear right from the start that the use of probabilistic models does not require that we assume that the data itself is in a linguistic sense variable, or in any sense fuzzy or unclear. I will come back to this point; it is certainly possible within a probabilistic framework to deal with data in which the judgments are non-categorical and in which a grammar predicts multiple possibilities. But in order to clarify the fundamental points, I will not assume that the data are anything except categorical and clear.

Assume most of what you normally assume about formal grammars: they specify an infinite set of linguistic representations, they characterize what is particular about particular languages, and at their most explicit they specify sequences of sounds as well as sequences of words. It is not altogether unreasonable, then, to say that a grammar essentially *is* a specification of sounds (or letters) particular to a language, plus a function that assigns to every

---

The conclusion I would draw from this small historical lesson is that Newton, the empiricist, was quite right in wanting to develop a tight, mathematical model for the data of his problem, right in wanting to dispense with occult, or hidden, properties, and right in being willing to posit them nonetheless if the overall coverage of the data and the simplicity of the resulting theory supported the decision.

In what follows, I will try to take this lesson from Newton the empiricist, and show how ideas from the 20th century could offer linguists a new perspective on what it is that they do.

sequence of sounds a real value: a non-negative value, with the characteristic that the sum of these values is 1.0. To make matters simpler for us, we will assume that we can adopt a universal set of symbols that can be used to describe all languages, and refer to that set as Σ. I do not really believe this is true, but it is much easier to express the ideas we are interested in here if we make this assumption.

A grammar, then, is a function g with the properties in (1).

$$
\begin{aligned}
g : \Sigma^* &\to [0,1] \\
\sum_{s \in \Sigma^*} g(s) &= 1
\end{aligned}
\tag{1}
$$

The grammar assigns a probability (necessarily non-negative, but not necessarily positive) to all strings of segments, and these sum to 1.

A theory of grammar is much the same, at a higher level of abstraction. It is a specification of the set $\mathcal{G}$ all possible grammars, along with a function that maps each grammar to a positive number (which we call its probability), and the sum of these values must be 1.0, as in (2).

$$
\begin{aligned}
\pi_G : \mathcal{G} &\to [0,1] \\
\sum_{g \in \mathcal{G}} \pi_G(g) &= 1
\end{aligned}
\tag{2}
$$

To make thing a bit more concrete, we can look ahead and see that the function π_G is closely related to grammar complexity: in particular, the complexity of a grammar g is $-log\,\pi_G(g)$; likewise, the function g is closely related to grammaticality; in particular, $-log\,g(s) + \sum_{w\,in\,s} log\,pr(w)$ is a measure of the ungrammaticality of s.

Our claim here is that, *as far as we can see at this point in time*, algorithmic complexity and data-probability under a hypothesis are enough to solve the problem, and it is in this light that this paper should be read. The task of the grammarian, given a corpus of data D, is to find the most probable grammar g, and g's probability, in this context, is directly proportional to its probability based on its algorithmic complexity, multiplied by the probability that it assigns to the data D.

4   Probabilistic grammars

In order to make this new empiricist interpretation work, we need to introduce the notion of probabilistic grammar, developed first by Solomonoff in the 1950s (see Solomonoff 1997 [21]). Curiously, the notion is not widely known or appreciated in mainstream linguistics, and my impression is that most linguists think that probabilistic models make vague, soft, or non-categorical predictions. This is false; probabilistic grammars can be put to those purposes, but we will not do so, and there is nothing about probabilistic grammars that requires one to make fuzzy predictions. Rather, what makes a model probabilistic is much more formal and mathematical (see, e.g., Goldsmith 2007 [11]).

Like virtually any other formal device, a probabilistic grammar specifies a universe of possible representations for the domain it treats; but in addition, a probabilistic model associates with each representation a non-negative number, its probability, and a strict condition is associated with these probabilities: the sum of the probabilities of all of the representations must be 1.0—neither more nor less. Informally speaking, a probabilistic grammar can be thought of as possessing an infinitely dividable substance, referred to as probability mass, and it doles it out to all of the representations it generates. The goal is to find a grammar that assigns as much of that probability mass as possible to the data that was actually seen. In a sense, this is the crucial difference between the empiricist (and probabilistic) approach and the generative approach: the empiricist, like the rationalist, wants and needs to generate an infinite class of representations, but the empiricist measures the adequacy of the grammar on the basis of how well the grammar treats data that was naturalistically encountered (that is to say, data that was recovered from Nature in an unbiased fashion).

The condition that the sum of the probabilities of all generated representations be equal to 1.0 is trivial in the case where there are a finite number of representations, to be sure. But it is typically not a problem when the representations form an infinite set either. If the reader is uncertain how it can be that an infinite set of positive numbers sum to 1.0, imagine that all the representations are sorted alphabetically, in such a way that shorter ones come first (that is, by treating space as the first element of the alphabet), and then assign probability 2^{-n} to the nth word. A moment's thought will convince the reader that these numbers sum to 1.0 ($\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + ...$)

To repeat, then: the first goal is to find the grammar that maximizes the probability of the observed data (this will be modified slightly, in due time, to allow the simplicity of the grammar to play a role in the selection). Any reasonable theory will assign most of the probability mass to unseen events,

that is to say, to sentences that have never been pronounced, and perhaps never will be. That's not a problem. The grammar will not be tested on the basis of those sentences, either: it will be tested on the basis of the probability that it assigns to the sentences that have already been seen.[10]

It should now be clear that the purpose of our insisting that a grammar be probabilistic has nothing to do with evaluating the probability of different sets of data. It would indeed be odd if we were to use a probabilistic grammar to decide what the probability was of various data that had in fact been observed. No; rather, the point of asking different grammars what probability they assign to a single, fixed set of data is to *evaluate the grammars, not the data*. If the data is naturalistic, then we know it exists, but what we care about is evaluating different candidate grammars to see how well they are able (so to speak) to decide which set of data actually exists, and we do this by seeing which grammar assigns the highest probability to the corpus.

We turn now to the subject of bayesian analysis. [11]

---

[10]Much of what I say here does not depend on that particular statement; one could adopt most of what we discuss and still believe that the heart of science is prediction, but I will not delve into this question.

[11]I recently noticed a comment made by Chomsky in 1968 [4] that made it clear that he was well aware of this issue, or so it seems to me. On pp. 76-77, he observes:

> A third task is that of determining just what it means for a hypothesis about the generative grammar of a language to be "consistent" with the data of sense. Notice that it is a great oversimplification to suppose thata child must discover a generative grammar that accounts for all the linguistic data that has been presented to him and that "projects" such data to an infinite range of potential sound-meaning relations....The third subtask, then, is to study what we might think of as the problem of "confirmation"—in this context, the problem of what relation must hold between a potential grammar and a set of data for this grammar to be confirmed as the actual theory of the language in question.

The problem that Chomsky alludes to here is exactly the problem that bayesian analyses aim to comes to grip with.

Suppose we take our goal to be to make explicit such a confirmation function $\mathcal{C}$ which maps from sentences to the positive reals, and let us make the following assumptions, which seem reasonable enough:

1. The degree of confirmation of a grammar by a set of data is inversely related to the value of the function;

2. The function is "extensive" at the sentence-level: that is, the degree of confirmation of a corpus which consists of two sentences, S_1 and S_2, is the sum of the confirmation of each sentence taken individually.

Then we can show that this confirmation function is equivalent to a probability function (or more accurately, to the negative logarithm of such a function). In particular, we use a

5   Bayesian analysis

A bayesian approach to probabilistic modeling is one that takes into consideration not only the probability that is assigned to the data by a model (or as we linguists say, by a grammar), but also the probability of the model (i.e., the grammar). And this latter notion is one that takes us right into the heart of classical generative grammar, to the notion of an evaluation metric. But first we will look at the mathematical side of Bayes's rule.

Bayes's rule involves inverting conditional probabilities, although from a mathematical point of view it is a very simple algebraic manipulation. What follows in this section is a bit dry, and the reader who wishes to may skip down to (7), and just take Bayes' rule on faith.

We need first to state what it means to speak of the probability of an event X, given another event Y, written $pr(X|Y)$. This means that we consider only those possible situations in which Y is true, and within that set of situations, we calculate what X's probability is. If we select a word at random from English, the probability will be about 8% that it is *"the"*, but if we look only at sentence-initial words, the probability of *"the"*, given that it occurs sentence-initially, is quite a bit higher. The probability that *"the"* occurs, given that it is in sentence-*final* position, is essentially nil.

To calculate such probabilities, when we already have in hand a system which assigns probability mass to all possible representations, we do the following. To determine the probability of X, given Y, we ask: how much probability mass altogether is assigned to all of the events in which both X and Y are true? And we divide this quantity by the probability mass that is assigned to all of the events in which Y is true. If we want to know the probability of the word *"the"* sentence-initially, then we calculate the probability that *"the"* occurs sentence-initially, and divide by the probability that a random word selected is sentence-initial. That is:

$$pr(X|Y) = \frac{pr(X \, and \, Y)}{pr(Y)} \tag{3}$$

But it will often be the case that we want to invert the dependence, in the following sense. We can calculate the probability that the word *"the"* occurs in sentence-initial position: that is the probability of *"the"*, given that it's in word-initial position. But we may also be interested in knowing, for any

---

Boltzmann distribution, in which the probability of a sentence s is $\dfrac{e^{-\mathcal{C}(s)}}{Z}$, where Z is the partition function, i.e., $\displaystyle\sum_{r} e^{-\mathcal{C}(r)}$, summing over all possible representations.

given occurrence of the word *"the"*, what the probability is that it is sentence-initial. If the first is $pr(T|I)$, then the second is $pr(I|T)$. Bayes's rule is the formula that relates these two quantities.

Expression (3) can be rewritten as (4), and a moment's thought shows that if we interchange the symbols *"X"* and *"Y"*, we obtain (5) as well.

$$pr(X|Y)pr(Y) = pr(X\,and\,Y) \tag{4}$$

$$pr(Y|X)pr(X) = pr(Y\,and\,X) = pr(X\,and\,Y) \tag{5}$$

And since the left-hand side of both (4) and (5) are equal to the same thing (that is, to $pr(X\,and\,Y)$), they are equal to each other:

$$pr(X|Y)pr(Y) = pr(Y|X)pr(X) \tag{6}$$

And then we have Bayes's rule, as in (7) .

$$pr(X|Y) = \frac{pr(Y|X)pr(X)}{pr(Y)} \tag{7}$$

Now, this rule is used in a very surprising way within what is known as bayesian analysis; we will take "X" to be a hypothesis H, and "Y" to be the set of observed data D. To make this more perspicuous, I will rewrite this and change the names of the variables:

$$pr(H|D) = \frac{pr(D|H)pr(H)}{pr(D)} \tag{8}$$

Now this says something much more remarkable from a scientist's point of view. Translating it into English, it says that the probability of a hypothesis, given what we have observed (and what else do we have other than what we have observed?) is equal to the product of two numbers, divided by a third number. It is the product of the probability that the hypothesis assigned to the data and the probability of the hypothesis in the abstract, divided by the probability of the observations themselves.

Suppose that's all true, and suppose that we can somehow come up with those values. It would then follow that we could choose our hypothesis out of a range of different hypotheses H by finding the one whose probability was greatest, given the observations. That's the heart of the notion of a bayesian analysis.

Of the three values that I described, only one is difficult to obtain, and that is the probability of the data, the denominator of (8). But we do not worry about that, because it does not really matter. Since what we care

about is choosing which hypothesis is the best, given the data, we are just going to keep $pr(D)$ fixed as we consider various different hypotheses. So the hypothesis h for which the value $pr(D|h)pr(h)$ is the greatest is the same as the hypothesis for which the value of $\frac{pr(D|h)pr(h))}{pr(D)}$ is the greatest, and that is the hypothesis we want. More mathematically, we say we want to identify the H as follows:

$$H = \operatorname*{argmax}_{h} pr(D|h)pr(h) \qquad (9)$$

This identifies H as being the hypothesis for which the product of the two probabilities defined there is the greatest. We still need to obtain two values: the probability of the data, given any of the hypotheses we are considering, and the probability of each of those hypotheses. We obtain the first by demanding that we only consider probabilistic grammars, which we introduced (following Solomonoff) in the previous section, and we obtain the second by establishing a prior probability over grammars. That is worth emphasizing: the H that we seek here is a generative grammar that assigns probabilities to its output. We will seek a way to distribute the probability mass over all grammars based just on what they look like as grammars, independent of how they treat any actual data. If we can do that, then the task of choosing a grammar, given a set of data, will be a matter of jointly considering two equally important things about the grammar: how good a job does it do of modeling the data, and how good is it as a grammar?

6   Establishing a prior probability for grammars

I am going to assume henceforth that the class of possible grammars is infinite. I don't think that there is a serious alternative to this hypothesis. Occasionally the suggestion is made that the real heart of a grammar of a human language is the correct selection of values assigned to a finite set of parameters (where each parameter can in principle only take on a finite number of values). But even if one believes in such a limitation (and as it happens, I do not), the "real heart" is only the heart: there's the rest of the grammar, which includes at the very least a lexicon, and I daresay no linguist would dream of saying that there is an upper bound on the size of a lexicon. The bigger the lexicon, the less likely it is, and its probability (to say nothing of its plausibility) shrinks very rapidly as its size increases.

Most theories of grammar are "non-parametric," in the specific sense now that grammars typically consist of formal (indeed, algebraic) objects which

can be made larger and larger, by adding more to them (even if the "more" is just another lexical item, or construction, phrase-structure rule, condition on a phrase-structure rule, etc.) What we do know about them, though, is that they are built up out of a specific set of formal objects, or symbols. There is no limit to the number of grammars, because there is no limit to the number of symbols (that is, number of occurrences of symbols) that may appear in a grammar.[12]

I would like now to be able to talk about the size or "length" of a grammar. Let's say that we build our grammars out of an alphabet of symbols $\mathcal{A}$, and there are S different symbols in $\mathcal{A}$. Let's also assume that it always makes sense to write out a grammar on a single line: we may prefer to draw things on a two-dimensional piece of paper, but for technical purposes, we will reduce all our notions to expressions written in one long string of symbols. It then follows that for any given length L, there are exactly S^L different string of symbols that could in principle be grammars. (Most of the strings will be formally meaningless in all likelihood, but that's OK, because we're trying to get an upper limit on things). For technical reasons that I will not go into,[13] we will assume that it is always possible to tell, from a purely formal point of view, when we have gotten to the end of the grammar (perhaps by setting up a symbol to specifically mark for that, or in any of a variety of ways).

We know one more thing about grammars that we want to use, and that is that a shorter grammar is always better than a longer grammar, all other things being equal. The reader may object to that, and say, "we've been there before, and done that, and don't want to do it again: sometimes the notation is doctored so that a shorter grammar is not the psychologically real one." To which I would reply two things: when we say "all other things being equal," we really and truly mean that we are making the claim that shorter is better only when we agree to fix and hold constant the theory of grammar; and second, we are not quite saying that *better = psychologically correct*. What we're saying is that if we are to assign a probability mass over an infinite class of grammars, then it must be the case that as we look at the class of longer and longer grammars (and they are vastly more numerous than shorter grammars, since for any length L there are S^L of them, and that expression grows quickly with L), the total probability mass assigned to them gets indefinitely small. For any amount of probability mass ϵ you

---

[12]The boundary between parametric and non-parametric analyses is getting a bit harder to draw these days. Goldwater's 2006 [13] employment of the Chinese Restaurant process blurs the line further, allowing most of lexicon generation to be viewed with a parametric model.

[13]This relates to the notion that our notation has the prefix condition, which relates in turn to satisfying the Kraft inequality.

choose, no matter how small, there is a length $\hat{L}$ such that the sum of the probabilities of all of the infinite number of grammars that are of length $\hat{L}$ (or greater) is less than ϵ .

There is one more crucial step to take, and that is one that permits us to escape from the clause that says, "given a fixed theory of grammar." Because we are not "given" a theory of grammar, after all; each of us is free to develop our own theory of grammar, and how can simplicity in my theory be compared with simplicity in your theory? What if my theory has (let's say) grammatical relations as a primitive notion, and yours doesn't? My theory allows me to write some grammars very simply that yours either can't express, or can only express with great complexity.

The answer I would like to suggest is based on algorithmic complexity (and thus is an application of ideas by Solomonoff, Chaitin, Kolmogorov, Rissanen, and, a little less directly, Turing; see Li and Vitányi 1997 [17] for details). The basic idea is this: any computation can be specified as a particular Turing machine, and there is, furthermore, such a thing as a universal Turing machine, and the latter is so important that we will give that phrase a three-letter abbrevation: UTM. Such a machine (and there are many of them) can be programmed to function like *any* other Turing machine, and in particular to accept programs in a higher level language, such as C, Lisp, or natural-language-grammar-language. If there were only one such machine, we could use the length of the program in its language as the basis for our notion of complexity, but the fact is that there are many, different UTMs, so our problem is how to deal with the nature of the differences among UTMs.

The reader has undoubtedly encountered the notion of a Turing machine: it is a finite-state device which is connected to an infinite tape, a tape which in turn is broken up into boxes in which only x's and blanks appear. The input to the machine is written by us in the first instance, and the machine can rewrite what it sees on the tape according to its internal program. Anyone who has actually looked at instructions to a Turing machine will be struck by how elementary the statements look: e.g., "If there is an x in the box you see now, erase the x and move one box to the right." But that's just typical of what instructions look like, even in real-world computers, at the level of machine language code. With real machines and also with Turing machines, one can enter a program written in a higher order language (like C or natural language grammar). In the case of a Turing machine, one does this by writing down two long things on the tape before beginning: the first is a compiler for the higher language (it is, so to speak, a program written in the UTM's native language which will input what follows it on the tape, view it as a program and translate it into the UTM's native language), and the

18

second is the program in the higher order language. If the Turing machine is truly a universal Turing machine, then it can be made to imitate any *other* Turing machine: that is, it's always possible to write a program which, if it is used to precede any chunk of data on the tape, will cause the universal Turing machine to treat that data like the Turing machine you wish it to imitate. (To put the same point slightly differently, there is a rough and ready equivalence between Turing machines and higher-level programming languages).

Given a particular universal Turing machine UTM_1, our job is to write a compiler which allows us to write natural language grammars. We will call such a compiler a $UG(UTM_1)$: it is a Universal Grammar of UTM_1. Our job as linguists is to write a $UG(UTM_1)$, and then to help other linguists write grammars for particular languages in the particular style that $UG(UTM_1)$ implements. Whenever we wish to implement a grammar for a language, we will give as input to the machine a string in two pieces: first, the $UG(UTM_1)$ and then the grammar written for it. In particular, we define a given UG by an interface, in the following sense—we need to do this in order to be able to speak naturally about one and the same UG being run on different UTMs (a point we will need to talk about in the next section). A UG specifies how grammars should be written, and exactly what it costs to write out any particular thing a grammarian might want to put into a grammar. Naturally, for a given UTM, there may be a large number of ways of implementing this, but we care only about the simplest one, and we will henceforth take it for granted that we can hire someone and outsource the problem of finding the implementation of a particular UG on any particular UTM.

Once we have at least one such grammar, we can make a long tape, consisting first of $UG(UTM_1)$, followed by a Grammar for English (or whatever language we're analyzing), as we have already noted—plus a string of words; the system will parse the words and produce as its output a labeled diagram of the sentence, and then halt.

Unfortunately, there will be many different ways of accomplishing this. Each UTM is consistent with an indefinitely large number of such universal grammars, so notationally we'll have to index them; we'll refer to different Universal Grammars for a given UTM (let's call it UTM_i) as $UG_1(UTM_i)$ and $UG_2(UTM_i)$, etc. Think of this as normal within linguistics: there are different theories of grammar, and each one can be thought of as a compiler for compiling a grammar into a machine-language program that can run on a UTM. A UG is intended to be used to write grammars for all languages of the world. At any given time (which is to say, at any given state of our collective knowledge of languages of the world), for any given UTM, there will be a best UG; it is the one for which the sum of the length of UG, plus

the sum of the lengths of each grammar written in UG, plus the compressed length of the data for each language in its corresponding grammar is the shortest.

We are almost finished. We now can assign a probability to a grammar that has been proposed. Given a universal Turing machine UTM_1, a univeral grammar UG_1 written for it, and a grammar G written for universal grammar UG_1, the probability assigned to it is

$$pr(G|UG_1) = \frac{1}{2^{length(UG_1)} S^{Length(G)}} \tag{10}$$

where S still indicates the number of symbols in our symbol system (and we assume that UTM_1, like all universal Turing machine programs, is written in binary code). In effect, this is the simplest way to divide the probability mass up over the entire universe of possible universal grammars and language-particular grammars, and it extracts the probability that goes to just this one (=this universal grammar and this grammar).[14] There is a lot of serious work that has gone into this equation, and I have only skimmed over the surface here—but bear in mind that this is directly responding to the needs of equation (9) above. We now have a prior probability distribution over grammars, which is what we needed by the end of the previous section in order to develop a Bayesian approach to linguistics and grammar justification. What we have not addressed yet is the question of uniqueness. Since there are many universal Turing machines, we could expect that the distribution over $(UTM_i + grammar_j)$ will vary with the choice of universal Turing machine.

This is an interesting question, to which I will give a sketch of an answer in the following section.

7   Your Universal Turing Machine or mine?

The case that would concern us is the case where the choice of two different universal Turing machines would lead us to select two different Universal Grammars (UGs).[15] What should we do if we find that at some particular point in our knowledge of languages, there is a UG, UG_1, which runs on UTM_1, and it outperforms every other UG on all the languages of the world.

---

[14]Again, for simplicity's sake, I am assuming that our UTM's can be extended to allow input strings which contain all the symbols of an alphabet such as $\mathcal{A}$.

[15]I am grateful to Mark Johnson for a conversation on this point, without which I would be even more confused than I probably am anyway. There is no shame in that admission; I am in good company.

But on UTM_2, UG_2 outperforms every other UG, including UG_1, on all the languages of the world. We can assume that there is at least one language for which the two UGs select different grammars for the same data; we will restrict our attention to the data from that language, and the grammars for that language.[16] What should we do?

Back to the problem of whose UTM we are going to use as our reference. Our problem case will arise as follows. Suppose we have data from one language, and two grammars, G_1 and G_2. If we choose UTM_1, then G_1 is preferred over G_2, while if we choose UTM_2, then G_2 is preferred over G_1.[17] This would happen if

$$|UG_1|_{UTM_1} + |G_1|_{UG_1} < |UG_2|_{UTM_1} + |G_2|_{UG_2} \tag{11}$$

but

$$|UG_1|_{UTM_2} + |G_1|_{UG_1} > |UG_2|_{UTM_2} + |G_2|_{UG_2} \tag{12}$$

(Imagine, if you'd like, that UTM_1 permits some important and complex operation to be expressed simply and this operation is used by G_1, but UTM_2 does not. However, except for that difference G_2 is a better grammar, i.e., shorter.) Now, because these UTMs are in fact universal, this means that there is a translation program from one to the other, and in fact for each pair of UTMs, there is a shortest translation device used to allow us to simulate UTM_i by using UTM_j; that is, we could say that it turns a UTM_j into a UTM_i. Let's indicate the length of the shortest such "emulator" as $[j \gg i]$,

---

[16]There is a slight irregularity in my doing this, which the careful reader will note. The differences between the ways that the two UGs work on all the other languages are being ignored, and that is not quite right. Again, we're trying to get to the heart of the matter.

(If you are following carefully, you will notice that it's not always obvious that we can talk about the *one and the same* grammar being run on two different UTMs each with their different UGs (that is, $UG(UTM_1)$ and $UG(UTM_2)$). What if one of the UGs allows us to refer to "subjecthood", for example, and the other UG has no way to talk about subjecthood at all? These concerns can make the whole problem very messy. Let's try to keep things simple, and for present purposes assume that any UG can in some obvious sense encode any grammar that another UG can, but the *length* (i.e., complexity) may vary greatly from one to the other. We will also assume that we can make sense out of the idea that one and the same Universal Grammar can appear in two different implementations for two different UTMs. That does not seem problematic, but I am assuming our ability to resolve a number of technical problems.)

[17]Let's assume for this first discussion that the two grammars G_1 and G_2 are equally good at assigning a probability to the data D. Obviously that will never be true, but if we didn't make that simplifying assumption, the discussion would be even more complex than it is, without changing the basic nature of the problem, which is about how universal any given UTM is or isn't.

which is necessarily greater than 0; think of it as "the size of a program that turns a j into an i". Then it follows that on UTM_1, UG_1's analysis of grammar G_2, using the best universal grammar it has access to, can never be longer than UTM_2's analysis of the data using grammar G_2 plus the cost of emulating UTM_2 on UTM_1, which is $[1 \gg 2]$. Informally speaking, a UTM will emulate another machine if the emulation does better than its own native performance, taking the cost of the emulator into account.(Warning: that may be a hard idea to visualize.)

If we translate this into inequalities, then we have the following.

$$[2 \gg 1] + |UG_2|_{UTM1} + |G_2|_{UG2} > [2 \gg 1] + |UG_1|_{UTM_1} + |G_1|_{UG_1} \tag{13}$$

$$[2 \gg 1] + |UG_1|_{UTM_1} + |G_1|_{UG_1} \geqq |UG_1|_{UTM_2} + |G_1|_{UG_1} \tag{14}$$

$$|UG_1|_{UTM_2} + |G_1|_{UG_1} > |UG_2|_{UTM_2} + |G_2|_{UG_2} \tag{15}$$

Putting these together, we get

$$[2 \gg 1] + |UG_2|_{UTM1} + |G_2|_{UG2} > |UG_2|_{UTM_2} + |G_2|_{UG_2} \tag{16}$$

Or

$$|UG_2|_{UTM_2} - |UG_2|_{UTM_1} < [2 \gg 1]. \tag{17}$$

By symmetric reasoning, we obtain:

$$|UG_2|_{UTM_1} - |UG_2|_{UTM_2} < [2 \gg 1]. \tag{18}$$

and

$$|UG_2|_{UTM_1} - |UG_2|_{UTM_2} < [1 \gg 2]. \tag{19}$$

What this says effectively is this: if you and I use different Universal Turing Machines to analyze a set of data, and my UTM is able to implement my grammar more easily (shorter) than it can your grammar; and if your UTM is able to implement your grammar more easily (shorter) than it can my grammar; then the discrepancy in the complexity of the theories of grammars used by my UTM and your UTM is bounded from above by the size of the emulators required by each of our machines to emulate the other.

More specifically, the difference between the complexity of the theory of grammar on my machine UTM_1 for my grammar (that's UG_1) and the complexity of the theory of grammar that your UTM_2 assigns to my grammar must be less than the cost of emulating my machine on yours. If you put it that way, it's obvious.

Now we have a choice of going in at least three different directions. First, we could say: fine; let's forget about choosing which UTM to use; let's just

pick one by convenience and start working. Second, we could say, how big do emulators tend to be? Are they dozens of bits long, or thousands, or billions, or what? If they are reasonably small, then these inequalities should stave off our concerns, and we can go back to choosing a random but reasonable UTM.

Or third, we could make a rational effort to find a means to select the UTM which was most universal, in the sense of minimizing the cost of translating from it into any other—that is, it would be the least biased. Given a set of UTMs $\mathcal{U}$, we select the member u for which this cost is a minimum.

Both the second and the third options have much to be said for them, and should be pursued, but in this paper I will pursue only the third.

We're nearly done. We can now compute (20):

$$pr(G|UG_1) = \frac{1}{2^{length(UG_1)} S^{Length(G)}} \tag{20}$$

This is the answer to the question as to how to find the best theory for a set of data D.

We are getting close to laying our hands on a trustworthy UTM. The problem was that some UTMs are secretly helping out some theories of grammar, by making them easier to write. We want to identify UTMs like that, and disallow them from being used to decide on what the best UG is. Our task now is to define a measure by which we can select the best UTM out of a given set of UTMs. We propose, as a natural candidate:

$$\operatorname*{argmin}_{u} \sum_{v \in \mathcal{U}} [v \gg u] \tag{21}$$

This says that we should select the UTM which can be translated into other UTMs with the fewest hiccups (i.e., with the least evidence that it contained unnecessary structure). How can we find this UTM?

It remains to be demonstrated that we cannot specify a feasible algorithm to find this UTM. But we do not need such an algorithm. We leave this step to the process of the advancement of science. Every researcher who subscribes to this practice has an interest in creating a UTM that will favor his or her approach as much as possible, and will (more importantly) not contain any hidden tricks that favor a competitor's approach. All that is necessary is an inventory of grammars of languages and of competing UTMs; it does not matter if they have been created automatically or by hand. The criterion in (21) will specify which is the right UTM for *all* researchers to employ: it is the one which is the most unbiased as to theory.

I will try to sketch some of the points discussed above a bit more graphically. Classical generative gramamr proposed that the problem of grammar
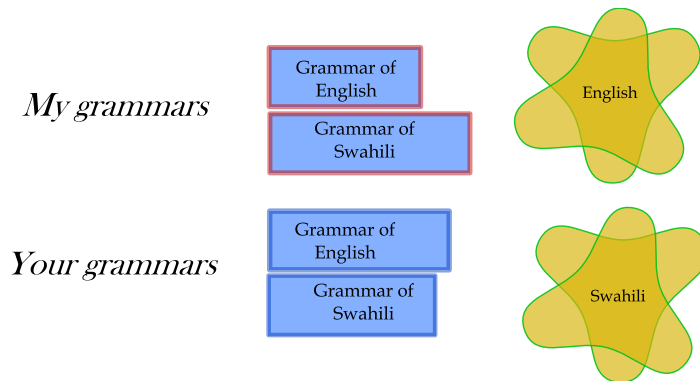
Figure 1: Classical problem

selection was that in Figure 1 (where the brown figure on the right are the corpora), we can't decide if my grammars are better than yours, even if each of us stick to a consistent pattern of grammar-writing: we have to take into consideration how much information is intended to be language-independent in our grammar-formulation. Each of us need to factor our grammars into a Universal part and a language-particular part, as in Figure 2. Each of us adds up the length of our three blue boxes, and whoever has the shortest sum (that is, blue area) wins.

The problem for generative grammar is two-fold: first, it does not take into account how well the grammars treat the data (that is the standard problem addressed by MDL and algorithmic complexity), and second, it overlooks the fact that we may get inconsistent rankings of the total area calculated depending on what system we use to program our Universal Grammars in. This is illustrated in the more complex Figure 3.

The solution that we propose is to *eliminate* the UG component, and select the best universal Turing machine on the basis of practice sketched above, which is roughly inspired by the "I cut, you choose" principles of equity that all children understand!

8   Stepping back from the precipice

Enough of the math. What's this all about?

In simplest terms, it's been a justification of an empiricist definition of linguistic analysis. It is empiricist primarily in the sense that it offers an account of a body of observations. Of course, one could take the resulting grammar and hypothesize that it was in some sense embodied in the brain, but that would have to be tested in brain-particular ways. It is a call to
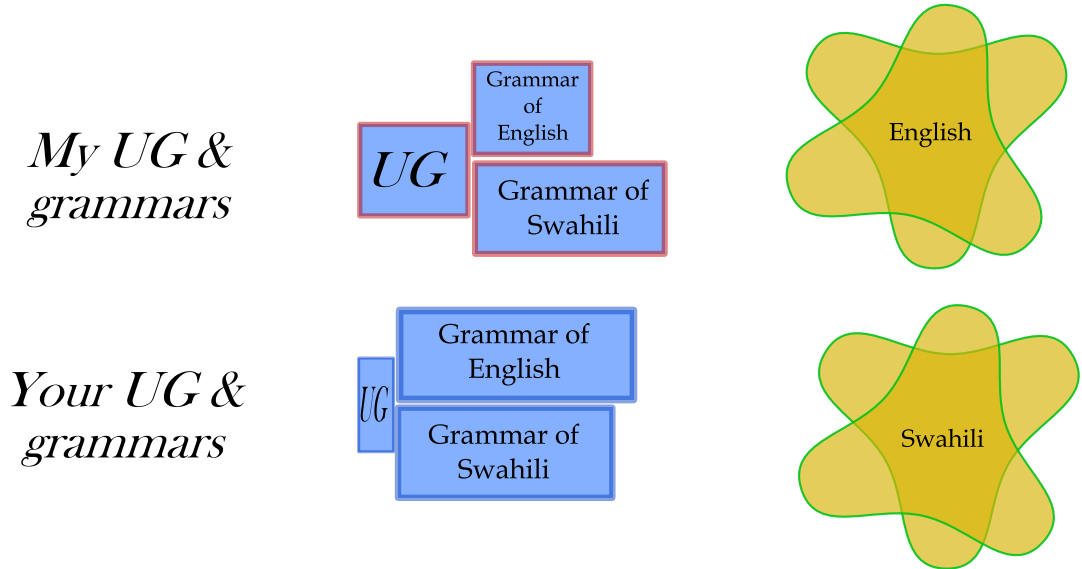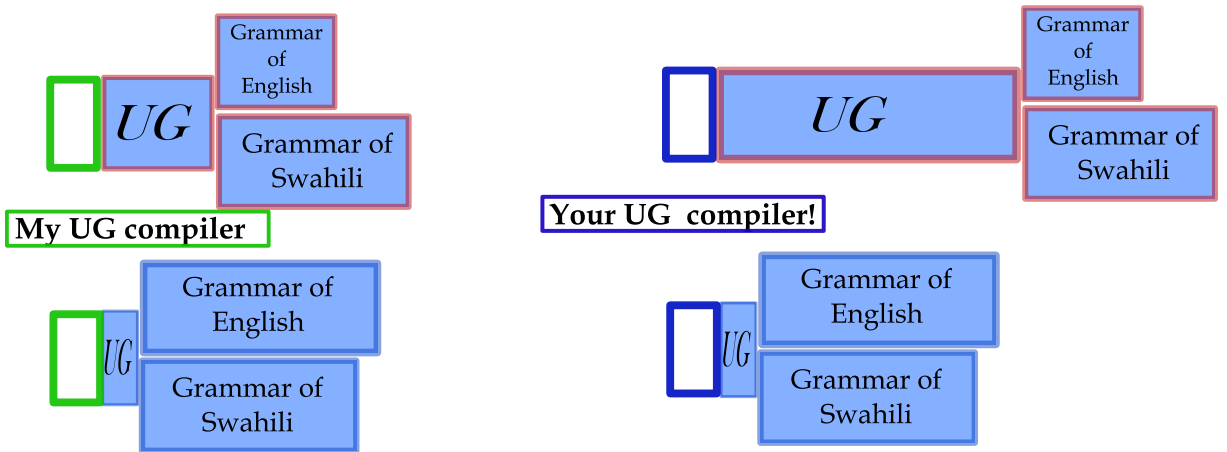
My UG &
grammars

Grammar
of
English

UG

Grammar of
Swahili

English

Your UG &
grammars

UG

Grammar of
English

Grammar of
Swahili

Swahili

Figure 2: Classical generative solution

Grammar
of
English

UG

Grammar of
Swahili

My UG compiler

UG

Grammar of
English

Grammar of
Swahili

Grammar
of
English

UG

Grammar of
Swahili

Your UG compiler!

UG

Grammar of
English

Grammar of
Swahili

Figure 3: Problem for classical solution

take both formalism and data very seriously—more seriously, I think it is fair to say, than it is being treated currently. I believe that someone should sit down, right now, and select a particular and convenient universal Turing machine, and write a compiler for a language in which grammars can be written. We can start small, with morphology and phrase structure rules; but let's get started, at least.

So much for the positive side. This approach may seem a bit heavy on the math side, but it seems very reasonable to me. The kind of linguistic work that it demands is not very different from what we are doing already, though there is some start-up cost associated with developing some software tools, so to speak.

This program may not suit many linguists. I recognize that it is easy to get to the point where one cannot understand how other people could hold differing views; the general outlines of what I have written here seem to me to be so sensible that at times I have to remind myself that not everyone views things this way. And yet it is true: many linguists do not want to think of themselves as empiricists. I am not quite sure why this is so, generally speaking, but there are a few indications present in the literature.

Noam Chomsky seems to be the most articulate spokesperson for the view that is down on empiricism, and as one of the many people who have learned a large part of what they know about linguistics from him, I for one think it is useful to hear what he has to say on the subject. I think he would be appalled at what I have written. But I do not agree with his view of the relationship of theory and evidence.[18] At times Chomsky has gone so far as to suggest that his method of science is one which allows a serious researcher to ignore data when it is incompatible with his theory. In a recent interview, Chomsky discussed both the methodological notions (which is what concerns us here) and some of the substantive notions involved in minimalism. Chomsky looked at Galileo, and said that

> [w]hat was striking about Galileo, and was considered very of-
> fensive at the time, was that he dismissed a lot of data; he was
> willing to say, "Look, if the data refute the theory, the data are
> probably wrong." And the data he threw out were not minor.

---

[18]If it wasn't obvious already, I might as well acknowledge that this paper is a dialog with a position that fits my best understanding of Chomsky. He has posed many of the right questions in a way that many others might have but didn't, and if he is often dismissive of others, I do not want to be dismissive of his views at all. (If you think "dismissive" a bit harsh, I would offer Chomsky's discussion of empiricism in [6] as an example of what I have in mind.) Which, needless to say, is not the same thing as saying I agree with him, or with them.

For example he was defending the Copernican thesis, but he was
unable to explain why bodies didn't fly off the earth; if the earth
is rotating why isn't everything flying off into space?...He was
subjected to considerable criticism at that time, in a sort of data-
oriented period, which happens to be our period for just about
every field except the core natural sciences. We're familiar with
the same criticism in linguistics....that's what science had to face
in its early stages and still has to face. But the Galilean style...is
the recognition that it is the abstract systems that you are con-
structing that are really the truth; the array of phenomena are
some distortion of the truth because of too many factors, all sorts
of things. And so, it often makes good sense to disregard phe-
nomena and search for principles that really seem to give some
deep insight into why some of them are that way, recognizing
that there are others that you can't pay attention to. Physicists,
for example, even today can't explain in detail how water flows
out of the faucet, or the structure of helium, or other things that
seem too complicated....the Galilean style referred to that major
change in the way of looking at the world: you're trying to un-
derstand how it works, not just describe a lot of phenomena, and
that's quite a shift. [1][19]

Chomsky summarizes the Galilean style as "the dedication to finding
understanding, not just coverage." Of course that sounds great—there is no
one who is *against* understanding, not among academic sorts who read and
write papers (like this one), and even pre-Galilean people were in favor of
understanding.

It's certainly a wildly inaccurate description of what Galileo was doing to
suggest that his methodological advance was to ignore data, and I find it hard
to conceive of why Chomsky would offer that interpretation, other than as a
justification for urging others to ignore data when the data contradict their
favorite theory.[20] If Galileo's insight was not to ignore data, then what was

---

[19]As I noted in footnote 3, one of modern empiricism's central concern was to clarify
what statements made by a science were empirical and which were conventional or logical.
This concern was perfectly legitimate. A scientific theory can seem to make claims, while
all the while having apparent falsifications be ruled out as irrelevant; such steps can be
taken by labeling certain phenomena as "stylistic" rather than grammatical, for example.
An overeagerness to hew to what Chomsky here calls the Galilean style—which is to say,
freedom to ignore data—can easily transform what was once a statement about reality
into a convention masquarading as science.

[20]As if it needed to be spelled out, the problem with this methodology is this: it works
just fine for me, but it is not fine for you, as far as I'm concerned. I am confident about

it? First of all, he came to his work with a deep and thorough skepticism regarding the received truth of the day, which was the Scholastic interpretation of Aristotle. In Chomsky's student days, the equivalent would have been a deep and thorough skepticism regarding American structuralism; in today's world, it would be a deep and thorough skepticism regarding minimalism.

Beyond skepticism, though, Galileo's theorizing was based on two principles before all others, and he could not have said any more clearly what they were: first, that we must look not to books but to Nature, the real phenomena, if we are to understand the world, and second, the language in which Nature is written is mathematical, which is to say, quantitative in character. It was not for nothing that Galileo is remembered for measuring the distance traveled by ball rolling down inclined planes: it was the study of what things really do that allowed him to show that these patterns did not fit the received wisdom of the time, no matter how well those theories satisfied the intellectual palettes of established scholars.

The fact is, there is no philosophy of science that allows one to ignore data. There is something else, though, which we can do when we see our theories running into empirical difficulties: we can acknowledge that our theories are still imperfect, and are inadequate for accounting for many things linguistic. There is no shame in that. There is nothing wrong with a science, such as linguistics, allowing for some research programs to be conducted despite poor empirical results, if there is enough agreement that the hypotheses may pan out someday; this is the scientific equivalent of the "let a thousand flowers bloom" philosophy.

There is a much deeper flaw, though, in Chomsky's logic, or perhaps it is his rhetoric. Remember: Galileo *wasn't* a success until his theories had been established empirically, both by matching prediction to observation, and by showing that apparent mis-predictions were only apparent and not real. There's no merit in ignoring data at the time; the only merit is in retrospect, after the stunning predictions actually do match the observations, when the scientist can pat himself or herself on the back for having never given up on a theory that eventually paid off. Holding on to a theory whoses predictions don't match the facts is like holding on to some stock in the stock market when everyone else says you should sell. You probably should sell, but if you don't, and you eventually make a million dollars from it, then you can tell everyone how smart you are. But you can't start telling them how smart you are until the stock actually goes up in value. There are far, far more

---

my own ability to identify true conjectures which do not appear to be supported by the data, but I am not so confident about yours. And, who knows?—you might feel exactly the same way about it, only in reverse. That's the problem.

people who have held onto to theories that never came back to life than there are people whose hunches overcame initial disappointment. It is romantic to think that holding on to a theory that seems to have been falsified is what made Einstein Einstein, but that kind of thinking won't work for cold fusion (or if you are still holding out for cold fusion, choose your favorite once-exciting-but-now-shown-false theory to make my point).

When all is said and done, it would verge on the irrational to deny that the long term goal of our research is to produce theories that simultaneously account for all of the relevant data, and do so with a minimum of assumptions.[21] The new empiricism offers a way to measure success along these lines. It may not be successful—we may find that probabilistic models cannot be established for some important areas, or that surprisingly arbitrary constraints need to be imposed upon the class of possible grammars. But it seems to me that we stand to learn a great deal from trying it out: we will learn where it succeeds, and I am sure we will also learn in the places where it may fail.

Thus Chomsky's first argument against this sort of empiricism may be summarized (not unfairly, I think) as: *we should follow the footsteps of the original scientific revolutionaries.* The response to this is that Chomsky has both misread the historical record, and failed to propose a methodological canon that we can all share (that is, it cannot be the case that we all get to choose which hypothesis is maintained regardless of the data; there will have to be shop stewards—or mandarins, or power brokers—who get to decide; I've tried to suggest that this is a hopeless and unattractive position to maintain).

Chomsky has offered a different argument, and one that carries more conviction, I think, but it too is based on just exactly what it is that we mean by *science.* In summary, his argument is that linguistics is either about something in the real world, or it is not. If it is about something in the real world, the only reasonable candidate about which linguistics can make claims is the human brain. If linguistics is not about the human brain, then it is not about anything in the real world, and there is therefore no truth of the matter, and therefore any linguist is free to believe anything s/he wishes to believe, and there are no scientific guidelines or standards—and in particular, linguistics is then not a science. Hence, if linguistics *can* be a science, then it *must* be a science of the brain.

That, of course, is my summary of Chomsky's idea, offered in a number of places. Here is one place, where Chomsky is responding to critics whose

---

[21]See the Afterword (page **??**, below) for some additional discussion of the role of data in scientific explanation.

criticism he finds impossible to fathom, linguists who do not believe that they are making claims about the human brain:

> Since there are no other objects in the natural world that the linguist's theory is about, the demand apparently is that the linguist construct a theory of some non-natural object. Again the tacit—and sometimes explicit—assumption seems to be that there are entities independent of what people are and what they do, and these objects are what theories of language are about, and further, must be about, on pain of irresponsibility. Again, we are left in the dark about these curious entities and how we are to identify their properties. Considerations of communication, the theory of meaning, the theory of knowledge, and folk psychology have also been adduced to argue that there are independent entities, external to the mind/brain, of which each of us has only a partial and partially erroneous grasp, always leaving as a mystery the manner in which they are identified, except by stipulation, and what empirical purpose is served by assuming their existence. I think there are ample grounds for skepticism about all of these moves...[22]
>
> Adopting this approach ,we abandon the hopeless search for order in the world of direct experience, and regard what can be observed as a means to gain access to the inner mechanisms of mind. (p. 19)

Before getting to the two specific issues at play here, I can't help remarking that it is pretty clear that Chomsky is incapable of putting himself in the shoes of the person who has a different perspective on this than he does (unless it's just a rhetorical stance—but I really don't think so). I say this because part of his extended discussion of this subject (and why not everyone agrees with him on these issues) includes the following digression:

> ...the shift in perspective was controversial at the time, and remains so today. One leading structuralist, Fred Householder, commented derisively (1966) that 'a linguist who could not devise a better grammar than is present in any speaker's brain ought to try another trade.'

---

[22]Chomsky 1997 [7], p. 18; all citations in this section are from that source.

Leaving aside some not unimportant considerations,[23] the heart of the matter comes down to two issues: *method* as an essential aspect of science and *physicalism*. The position that I defend is that method matters, and that method is a part (though only a part) of what constitutes science; and a denial of physicalism (the view that what can be said to exist must be identifiable, and specifically locatable, in space and time). I will only mention the first point briefly: a brief paper cannot do the point justice, and a longer discussion can be found elsewhere (see [12]). Oversimplifying the point, it boils down to this: everybody wants to do a kind of linguistics that merits being called a science. But to decide whether a particular style of linguistics merits that label, we need an account of what science is. There are several accounts we can turn to, but it is very important to bear in mind that dif-

---

[23]Two comments on this remark of Chomsky's: the first is its characterization of Fred Householder as a leading structuralist. I was a colleague of Householder's for eight years; he interviewed me and hired me for my first job, and I knew his views pretty well. He had enormous respect for generative grammar; he used it, he taught it, he published analyses within its perimeter, and he hired students of Chomsky's for his department. It's not clear to me that there are any other useful measures by which he would not be viewed as a generativist. The only respect in which he differed from other generativists was that he knew that there was a world of linguistics that existed before generative grammar, and he raised questions, in print, that he thought needed to be answered by generative grammarians, notably in his 1965 paper [14]. Fred was no more a structuralist than he was a generativist: mainly, he was just a linguist. I don't think Fred would have *objected* to being called a structuralist, but he would have been a bit puzzled by the observation. For my part, I must confess that I read Chomsky's characterization as a charge, a bit like Lakoff's famous example along the lines: "Noam called Fred a structuralist, and then *he* insulted *him*." Fred wrote once, in precisely this context: "Nothing can be so clearly and carefully expressed that it cannot be utterly misinterpreted." In the obituary for Householder in *Language* 73(3): 560-570, he is quoted as saying that in the 1950s, "Edward Stankiewicz lent me the first three chapters of the *Logical Structure of Linguistic Theory* by Chomsky. That really got me excited."

The second respect in which the remark is odd is that Chomsky does not seem to *get* Householder's point, to understand it—but Householder's point, I must say, is not very obscure, and it is certainly not 'derisive'. His point was that given a finite amount of data, a clever enough graduate student will often be able to come up with an account that is formal simple and yet for all that linguistically wrong. As Morris Halle famously put it in his phonology classes: *not every good idea is a right idea*. A phonological analysis that reconstructs proto-Indo-European forms, and then posits an order set of rules that exactly matches the sound changes that have occurred since, is one that could easily fall under the Householder's description: the analysis might be neat and tidy, but it is utterly underwhelming to the professional linguist. Indeed, if one chooses to understand Fred Householder's point, it is clear that he was (as he always proclaimed himself to be, whenever given the opportunity to make the point) a believer in the existence of *right* and *wrong* in linguistic analysis. That's what Householder meant, and he meant that when one linguist presents another linguist with an analysis, it is reasonable to ask for justification beyond the fact that the account handles a restricted set of data.

ferent accounts are motivated by different aims. The most famous, that of Karl Popper ([18]), was motivated by a desire to keep certain disciplines *out* (notably Marxism and Freudianism), while some contemporary approaches are arguably motivated by the desire to show that there is no, and can be no, clear distinction between science and non-science. Some linguists turn to other sciences to learn methods that have proven effective in a neighboring field; others do it to demonstrate that methods that we have, or social structures (like degree-granting, or journal-refereeing) that we have, resemble those of the traditional sciences more than they resemble other disciplines. Scholars who emphasize methods in science have tended to value such characteristics as clear statement of hypotheses, the repeatability of observations, the possibility of utilizing mathematical techniques in a non-trivial way, the specification of methods to ensure against contamination of the materials studied, and so on and so forth. Such a list typically does not, and would not, include conditions on the belief systems of the scientists involved: it really does not matter whether they believe that the physical constants of the universe are what they are because God made them that way, or they are what they are for no particular reason. There is no point in being a member of a scientific community in which one must hold a particular philosophical belief; the most that we can insist on, from our colleagues, is that they and we satisfy conditions of scientific behavior. This is not behaviorism, obviously; it is just to say that doing science right means understanding what the method of science is.

The second assumption that Chomsky's argument rests on is *physicalism*, that is, the philosophical view that for any given scientific hypothesis, we must be able to identify a specifiable place, in space and time, about which we are talking: if there is a *truth* about language, it is a truth about things going on in everybody's brain.

The problem with Chomsky's invocation of this principle is that he uses it as a rhetorical rather than a scientific arm, in the following sense: like every other linguist, Chomsky and his colleagues are not capable of establishing where and how their theories of grammar are instantiated in the brain; but they interpret their failure to do so not as a challenge to physicalism, but as an I.O.U., that is, a promise that at some unspecified future date, a physical location in the brain will be found. *There is no difference between promising, some day, to do something in a yet undiscovered way and not promising to do it at all.* It is not the linguist's job to determine how the brain works: that is a good thing, since there are few linguists with any serious training in neuroanatomy. It is the linguist's job to figure out how language works, and as that challenge continues to be handled, linguists and neuroscientists will be able in the future to come up with a synthetic view.

Most importantly, in my opinion, is that the bayesian approach that we have discussed—which is, as I have suggested at several points along the way, close in spirit to *The Logical Structure of Linguistic Theory* (Chomsky 1975[1955] [5])—provides an answer to the challenge: how can there be a truth or falsity about linguistic claims if linguistics is not a science about what happens in the brain?

The answer is this: the goal of the linguist is to provide the most compact over-all description of all of the linguistic data that exists at present: let's call that data "the Internet", to make it slightly more concrete. We want to find the system of UG, as above—that is, the system that compiles a grammar into the machine code of a fair and equitable Universal Turing Machine—for which the sum of the grammar lengths, and the inverse log probability of the data, under the grammars, is a minimum. Since all of these terms are by definition positive, we know that there is a lower bound,[24] and hence there is a *correct* answer to the question: what is that UG?

9   So this is *empiricism*?

Having now gone through a particular account of scientific knowledge concerning language, let us touch on the question of why we should call this account *empiricist*. On one account, empiricism is the view that all knowledge comes from the senses, and yet we have put a lot of effort into building up knowledge of a priori simplicity based on Kolmogorov simplicity. This is empiricism? This is *just knowledge that comes from the senses*?

No. The empiricist I have described in this book is not one who thinks that knowledge begins and ends with what the senses provide. Is it fair to call someone like that an empiricist—someone who cares about theory above and beyond the description of the data?

One reason for calling this approach empiricist derives not so much from what it *excludes*, but on what it *includes*. The empiricist believes that a thorough and quantitative consideration of the fit of theory to data is as important as theoretical conciseness, and that theoretical insight must set shoulder-to-shoulder with empirical covereage in science.

Bas van Fraassen has presented a contemporary take on empiricism—presenting a form of empiricism that is markedly different from logical empiricism. For van Fraassen, the essential point is that while the empiricist may recognize that particular theories make claims about unobservable enti-

---

[24]Of course, we need to do some work to show that there there is no failure of compactness in the space of Universal Grammars, but that is a minor technicality, and quite irrelevant to the main point.

ties, the empiricist knows (or believes) that what the theory says about the observables is the only thing that justifies belief in any given theory. [say more]

Mario Bunge [3] has offered a critique of modern empiricism in the context of the social sciences that misses the point, it seems to me. But reviewing what he says can help clarify the issues. He begins with a distinction between *ontological* and *epistemological* empiricism, similar to the distinction we made above (ref.). He takes ontological empiricism to be the position that only experiences exist, and the epistemological version to be that only experience can serve as the source and test of ideas. (311). Bunge criticizes empiricism for its narrowness (p. 306), but his own characterization of empiricism is too narrow to serve as anything but a whipping boy; indeed, he sees only naive empiricists and logical empiricists (311), and has nothing good to say about either. In fact, naive empiricism is "untutored," and not good enough for "higher animals," which includes human beings! Any idea of some complexity, going beyond what is immediately perceived, is for Bunge evidence of the inadequacy of naive empiricism. (Bunge goes so far as to put the views of sociologists of science like Barnes, Bloor, and Latour in the same bag as common sense, and when they argue that there is no difference in kind between everyday activity and scientific activity, he chalks up this failure to their empiricism: their view is an "oversocialized version of Bacon's naive empiricism." (311) This is an unfair jab at both empiricism and common sense.)

The perspective I've described here founds scientific belief on an expression in two terms, one of which evaluates theoretical elegance and parsimony, and the other fit to facts. The empiricist knows there's a trade-off between the two, and hopes he has found the right trade-off. He values theoretical simplificity, and knows that it is every bit as important as good description of observation. But he does not confuse the two: postulating an entity in his theory allows for simplicity in the theory expression; it is not an additional obsevation.

The naive realist[25] thinks that the empiricist is deluding himself; the naive realist thinks that the empiricist really *does* believe that the objects described by the theory do exist, and that is why even the empiricist wants

---

[25]Not everyone who uses the term "naive" takes the trouble to define it, which leaves a reader's mind willing to read too much or too little into it. I will say what I mean by this characterization. I think that a naive defender of a position is one who thinks that his opponents disagree with him because they have not even considered the evident, obvious reasons for his position. In short, the naive realist does not realize that there are intellectually valid and defensible positions different than his own, and that empiricists say disagreeable things because they haven't thought much about the problem.

to scour the universe to discover whether neutrinos and gravitons and Higgs bosons and all sorts of exotic theoretical objects exist. The naive realist tells the empiricist: you see! you really do want to test whether the things that you postulate exist. If you can't find them, you (or your colleagues) will no longer be satisfied with your theory. Just accounting for the observations isn't enough, if you know that the entities you postulate to account for those observations cannot be found when the tests are run.

The empiricist thinks that the naive realist gets carried away with his big ideas. He can love an elegant theory as much as any person, but he knows that history has shown that dragging the entities postulated by a theory into the world of what is observed is difficult and treacherous. It's not difficult the way it is difficult to drag up a ship that has sunk to the bottom of the sea; it is difficult the way it is to get a stain out of a rug, it is difficult the way it is to find an equitable way to share the wealth of our society. These are difficult things to do, and even if we try hard we may not succeed in accomplishing these tasks, and it does not mean that we were misguided by trying to accomplish them. But our best hopes may not be met. That is how the empiricist feels about theoretical entities: it is very reasonable to undertake to find them in space and time, if they are entities that are postulated to exist in space and time. If we find them, we will have enhanced our observed universe. But science continues, and it continues in its full-bore fashion with no problems at all, thank you very much, whether those theoretical entities are observed, or observable, or neither.

The realist replies: if you do not believe that those entities really exist, then you cannot believe, really and truly believe, the scientific theory, and you do not have a real explanation for the regularities you have found in the data and the observables until you acknowledge that you also believe that the unobserved things the theory postulates do in fact exist, so that they can actually *cause* the things observed to come into existence.

The empiricist realizes that the stakes have grown larger now. The realist has just accused him of bad faith, deep down inside. But there are perfectly valid responses—at least two that this writer is entirely comfortable with. The first amounts to throwing the accusation back in the face of the realist, based on the law of rejected entities (a law which I have just invited, and which I will explain in a moment). The second retort is one that most modern scientists will not be comfortable with, so I will not pursue it at length, but I will mention it at least briefly, and it involves the status of mathematical objects. But first, the law of rejected entities.

The empiricist's reply to the naive realist is this: you did not really believe that your unobserved entities existed. If you want to insist that you really do believe they exist, you are going to have to acknowledge that it is

in a rather different sort of way than the way in which you believe that the Empire State Building exists, or the Pentagon, or your left foot. Because we all know that scientific theories change, and what was once strong motivation for believing that something (like caloric or phlogiston, or the ether of which electro-magnetism is but a vibration, or the passive transformation if you are a generative grammarian) may tomorrow melt like the morning dew. You, naive realist, you are willing to change your view as to what exists on the basis of learning a better theory! The very foundation of your belief in theoretical entities is the conciseness of the theory that links the theoretical entities to the observations, and if there is a better way to account for the observations, you have no trouble at all dropping yesterday's belief in phlogiston or grammatical transformations. You would never do that in the case of a physical object: if the World Trade Center is gone today, it is because something happened to it, not because we have a better theory today that does not need it anymore (whatever that might mean!). In short, the law of rejected entities: both realists and empiricists say good-bye to rejected theoretical entities with no trouble at all.

Bunge sees an argument for at least closet realism in the efforts of scientists (including those who may call themselves positivists) in the constant efforts of scientists to describe their observations in terms of theoretical and unobserved entities. Why does he bother to go past the simple cataloging of his own observations, and those of other scientists? If there is no truth because there is no assumption that the theoretical entities really exist, then why put hypotheses about them to the test? (p. 357) Alas, these questions reveal that Bunge fails to take the empiricist at his word: the empiricist demands an elegant theory that deals with the observations as well as possible. The observations are what they are, and the responsibility does not go away to respect them—and the elegance of the theory is of utmost importance, so much so that we will make great efforts to formalize the vocabulary so that we are reasonably sure that we are doing a good job of it. It is not fluff, not window-dressing: theoretical parsimony is deadly serious business.

> By denying the autonomous existence of the external world, or at least the possibility of knowing it objectively, anti-realism discourages its scientific exploration and its rational control. Shorter: Anti-realism is anti-scientific and impractical. By contrast, realism is not just one more philsophical extravagance: it is the epistemology tacitly inherent in factual science and technology. It inheres in the former, because the declared aim of scientific research is to explore and understand reality. And realism inheres in technology, because the latter's job is to design or redesign

36

feasible artifacts or plans aimed at altering reality. p. 358

10   Conclusion

The observations I've made in this paper undoubtedly sound quite abstract, but they do have a very practical side to them. The discussion here has been based on Minimum Description Length (MDL) analysis (Rissanen 1989 [20]), and MDL analysis lies at the heart of an automatic morphology learner that I have been developing for about ten years (see [10] and *http://linguistica. uchicago.edu*), and it and related approaches are being pursued by a number of researchers at this point.[26] I believe that the approach discussed here can be applied quite directly to linguistic problems, and the reader is welcome to see an example of that at the website indicated. In fact, the order of things has been quite the opposite of what might appear to be the case, given the present paper: in actual fact, the concrete applications came first, and the theory came later (which is the usual order, in the lives of theories). The goal is to develop an understanding of what it means to develop a grammar of a set of data which is explicit enough that it embodies the considerations that a human linguist applies in determining what is the best grammar for the data.

The hypothesis that I have made in this paper is very simple: that a universal measure of algorithmic complexity is enough to provide an explanation for properties of grammars. This may not be true, from the point of view of neuroscientists: it may be that it is necessary to define, for example, a highly restricted subclass of grammars that are possible human grammars, because we discover that the algorithmically simpler ways of accounting for the data in these languages is not the way used by the human brain. I don't think that there is any reason for such pessimism at this point, but it is certainly possible in principle.

But the main take-home point is that algorithmic complexity, working together with probabilistic grammars, allows for a very appealing conception of what linguistics is, and developing an empiricist conception of the task that is remarkably true to the spirit of Chomsky's *Logical Structure of Linguistic Theory*.

---

[26]Any list that I could provide of such work would inevitably leave scholars out who deserve to be cited, and so I will abstain from providing such a list; an internet seach on the terms *"unsupervised learning grammar"* will generate a long list of candidate resources in this area. Add reference to [16].

11   Acknowledgments

References

[1] Adriana Belletti and Luigi Rizzi. An interview on minimalism with Noam Chomsky. University of Siena, Nov 8-9, 1999 (rev: March 16, 2000), 1999.

[2] Harry Bracken. *Mind and Language: Essays on Descartes and Chomsky.* Foris (Dordrecht), 1984.

[3] Mario Bunge. *Finding Philosophy in Social Science.* 1996.

[4] Noam Chomsky. *Language and mind.* Harcourt Brace Jovanovich, New York,, 1972.

[5] Noam Chomsky. *The Logical Structure of Linguistic Theory.* Plenum, New York, 1975[1955].

[6] Noam Chomsky. *Language and Responsibility.* Pantheon, New York, 1979.

[7] Noam Chomsky. Language and cognition. In David Johnson and Christina Erneling, editors, *The Future of the Cognitive Revolution.* Oxford, 1997.

[8] Carl de Marcken. *Unsupervised Language Acquisition.* PhD thesis, MIT, 1996.

[9] Pierre Duhem. *L'Evolution de la mécanique.* Vrin, Paris, 1992[1903].

[10] John A. Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

[11] John A. Goldsmith. Probability for linguistics. *Mathématiques et Sciences Humaines*, 180(4):5–40, 2007.

[12] John A. Goldsmith and Bernard Laks. *Battle in the Mind Fields*. 2007.

[13] Sharon Goldwater. *Nonparametric Bayesian Models of Lexical Acquisition*. PhD thesis, Brown University, 2006.

[14] Fred Householder. On some recent claims in phonological theory. *Journal of Linguistics*, 1:13–34, 1965.

[15] Jan Koster. Linguistics, historicism, and the humanities. In *Reflections on Language and Language Learning: In honour of Arthur van Essen*, pages 3–19. John Benjamins, Amsterdam, 2001.

[16] Shalom Lappin and Stuart M. Shieber. Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, 43:1–34, 2007.

[17] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, Berlin, 1997.

[18] Karl Popper. *Conjectures and Refutations*. 1953.

[19] Willard van Orman Quine. Two dogmas of empiricism. *Philosophical Review*, 60:20–43, 1951.

[20] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.

[21] Ray Solomonoff. The discovery of algorithmic probability. *JCSS*, 55(1):73–88, 1997.