

Grammar as Science



Richard K. Larson

illustrations by Kimiko Ryokai

The study of grammar once enjoyed a central place in education, one going back to the classic liberal arts curriculum of the late Middle Ages. Grammar was, along with logic and rhetoric, one of the subjects in the trivium: the core group in the seven arts students were expected to master. The importance of the “big three” is reflected in our modern word *trivial*, which originally applied to knowledge regarded as so basic that it required no argument. Any educated person could be assumed to know it.

In an earlier time, studying grammar primarily meant studying Latin and Greek. Access to the classical languages meant access to the root cultures of the West, their literature and science. Latin and Greek were viewed as “special languages”: models of clarity, logical organization, intellectual subtlety, and economy of expression. Studying how these languages worked was viewed as something very close to studying the principles of logical, coherent thought itself. When other languages were analyzed, they were always analyzed on the model of Latin or Greek.

The curriculum in which grammar held its place of honor is obsolete now; the time when educated people could attend only to the classics of the West is long past. Furthermore, we now know that Latin and Greek are, by any reasonable standard, typical human languages: in no way clearer, subtler, or more logical than, say, Greenlandic Eskimo or Chinese. The old rationales for studying grammar are gone. Is the relevance of grammar behind us, too?

Not at all! In the last five decades, the subject of grammar has been reborn in a very different setting. Grammar has emerged as part of a new science, linguistics, that poses and investigates its own unique and fascinating set of questions, pursuing them with the same rigorous methodology found elsewhere in the study of natural phenomena. This new scientific perspective on grammar owes much to the linguist Noam Chomsky, who introduced it in the mid-1950s and who has contributed centrally to its development ever since.



When we study human language, we are approaching what some might call the “human essence,” the distinctive qualities of mind that are, so far as we know, unique to man, and that are inseparable from any critical phase of human existence, personal or social. Hence the fascination of this study, and, no less, its frustration.

—*Language and Mind*, p. 100

Noam Chomsky
Institute Professor
Massachusetts Institute of Technology

The idea of a “scientific” approach to grammar might strike you as odd at first. When we think of “science,” we usually think in these terms (see Goldstein and Goldstein 1984):

- Science is a search for understanding.
- Achieving understanding means discovering general laws and principles.
- Scientific laws and principles can be tested experimentally.

How do such notions apply to grammar? What is there to *understand* about grammar? What would general laws and principles of grammar be? And how might we test laws and principles of grammar experimentally, assuming we could find them in the first place? Our puzzlement about these questions suggests a certain implicit view of language, and the kind of object it is.

Language as a Natural Object

From a very early age, children appear to be attuned to the distinction between **natural objects** and **artifacts**. In an interesting series of experiments, psychologist Frank Keil has shown that whereas very young children judge the identity of objects largely on the basis of superficial features, at some point they begin to realize that certain kinds of objects have an inner essence that may sometimes be hidden or obscured (see Keil 1986). For example, before a certain age children will identify a black cat that has been painted to look like a skunk as a skunk, whereas after this age they identify a black cat painted to look like a skunk as a painted cat and not as a skunk. They realize that being a skunk involves more than looking like a skunk; the true identity of an object may be concealed by appearances.

Interestingly, in making this transition, children seem to draw an important distinction between natural objects, like cats and skunks, and artifacts (things made by humans). Although they judge a painted cat to be a cat nonetheless, they understand that an old coffeepot that has been modified into a birdfeeder is now really a birdfeeder. In other words, they see natural objects as having their own defining properties, whereas artifacts are whatever we make them to be, as a matter of convention.

Human language can be viewed in both these ways, as artifact or as natural object; and how we view it strongly shapes our reaction to the facts it presents us with. Language has been seen by many people as an aspect of culture, similar to other basic human institutions and traditions like tool-making or agriculture. In this view, languages are the product of human imagination and development: created by humans, taught by humans, and learned by humans. They are cultural artifacts possessing the properties and obeying the rules that we bestow on them, and the patterns or regularities we find in them are basically just matters of convention. Like the birdfeeder, language is what we've made it to be, and there is no more to say. There is no question of understanding anything, or discovering anything, or testing anything. It is this broad view of language, I believe, that leads to puzzlement when we think about grammar as science.

But language can instead be seen as a part of the natural world. In a series of influential works, Noam Chomsky has argued that human language is more correctly viewed as a natural object, analogous to a limb or a bodily organ (see Chomsky 2000a). True, language arose in the course of human prehistory, but it was no more invented or developed by humans than arms or lungs. Rather, language ability evolved, like other species-specific properties. Likewise, although languages develop in the course of human ontogeny, they are neither taught to nor learned by children, any more than children are taught to grow arms or learn to have hearts. Rather, we humans speak and in so doing provide the environment—the “nutrition,” to use a Chomskyan metaphor—in which language can grow and develop in our children.

Under this perspective, languages become objects of the natural world much like quasars or spinach leaves. They are entities whose properties and structure are to be determined by naturalistic investigation. Accordingly, when we are faced with a certain pattern or regularity in linguistic facts, we do not put it aside as a matter of convention; rather, we start to look for a “law” or principle that predicts the pattern and suggests an explanation. And we realize that the explanation may well be hidden to us, and need to be tested for experimentally. Adopting the naturalistic perspective opens up human language as a new domain, a fresh territory for scientific exploration.

The Terrain Ahead

This book is an introduction to the modern subject of grammar (now called **syntax**) from the perspective of language as a natural object. Its goals are twofold:

- To systematically explore some of the ideas and results in the new territory of syntax, and
- To provide experience with rigorous scientific reasoning and argumentation, and the development of scientific theorizing.

Successful exploration requires open eyes and a clear head. You need to be observant about your immediate surroundings (so you won't miss anything). You need to be mindful of how you got there (in case you need to retrace your steps or reconstruct your route for others). And you need to be logical about where you will go next (so you don't just blunder about).

This book consists of short units that usually involve some specific factual point(s) and a small number of ideas or concepts. These will be your "immediate surroundings" as we proceed. Try to read and master each unit in a single sitting. Be observant, and try to see all there is to see.

When the terrain is unfamiliar, where you are and how you got there are sometimes difficult to keep in your head. Maps are useful for this purpose. The units of this book are grouped into parts that form the map of the territory we'll be exploring:

- Meeting the subject and discovering its questions (Part I)
- Constructing a theory that attempts to answer the questions (Part II)
- Choosing between competing theories (Part III)
- Arguing for one theory versus another (Part IV)
- Searching for deeper explanation (Part V)
- Following the many consequences of a theory (Part VI)
- Enlarging and constraining the tools that a theory employs (Part VII)

Since these divisions mark the stages that researchers typically pass through in constructing a scientific theory in any domain, they make a good general "route plan" for us. At the beginning of each part, we will stop and do a "map check" to make sure we know where we've gotten to and where we should go next. Often we will consult a guide, someone more familiar with the area.

Science is tentative, exploratory,
questioning, largely learned by doing!
—“Rationality/Science,” p. 91



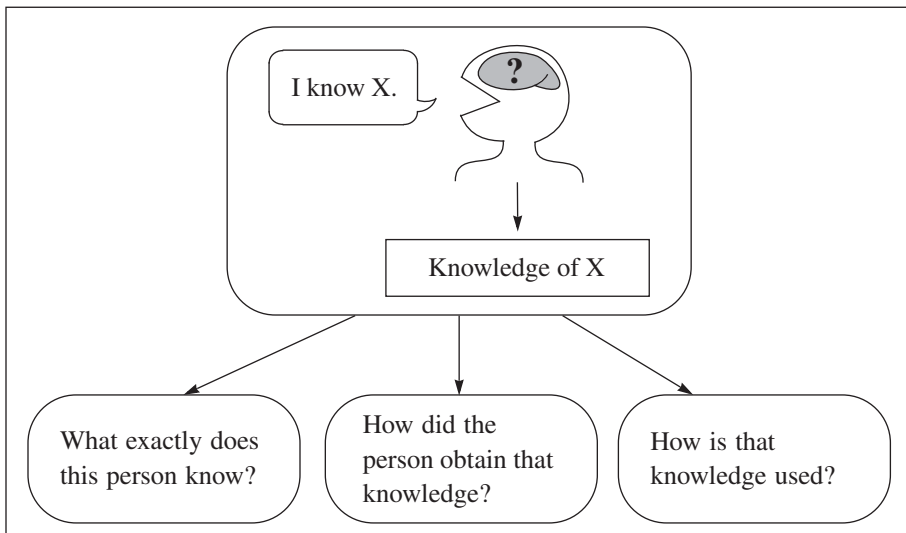
You won't need much in the way of equipment to undertake this trip. The presentation assumes no previous experience either with grammar or with the broader discipline of linguistics. All you will need is a healthy sense of curiosity and a willingness to think critically about a subject matter (language) that most of us take for granted in day-to-day life and rarely think about at all. With that much, we can begin.

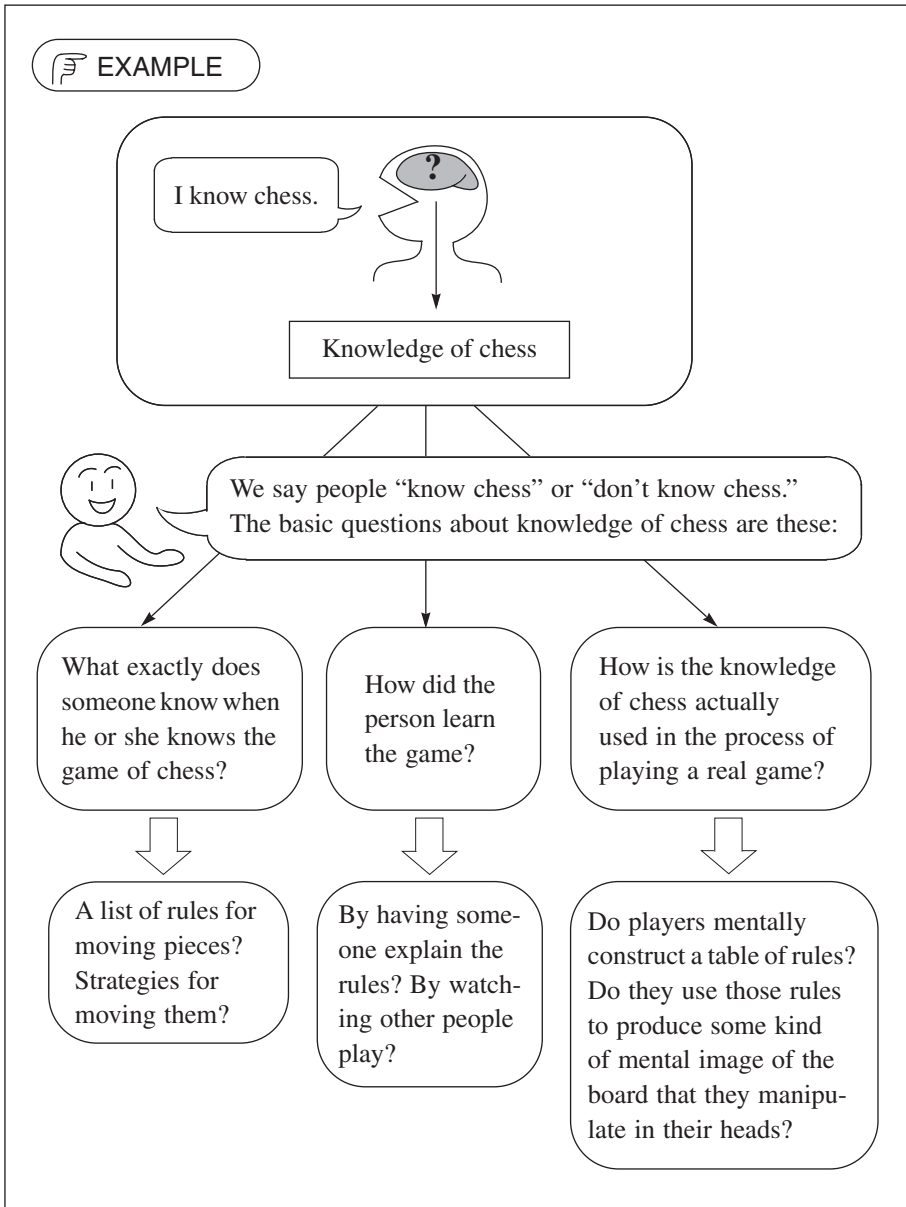
Leading Questions

In beginning the study of any field, one good way of orienting yourself is to find out what problems the field works on. What **leading questions** does it seek to answer? In the approach to linguistics we will follow, the leading questions are very easy to formulate.

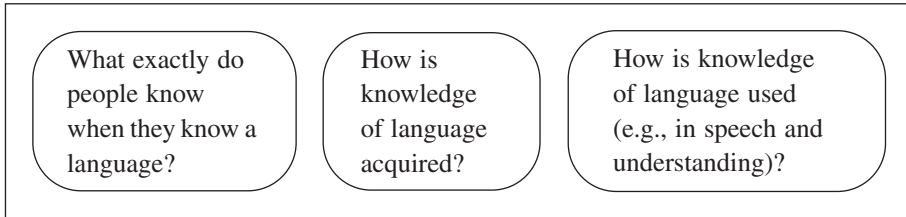
In day-to-day conversation, we routinely speak of people “knowing English” or “knowing Japanese and Korean.” We talk about a language as a body of knowledge that people do or do not possess. The leading questions of linguistics arrange themselves around this commonplace way of talking: they address **knowledge of language**.

Whenever someone can be said to know something, a number of basic questions present themselves.





Linguistics is concerned with these basic questions as they apply to knowledge of language. It seeks to discover the answers to these questions:



Viewed in this way—as addressing certain knowledge that we have internalized in the course of growing up—linguistics is basically a branch of **psychology**, broadly understood. Linguistics is trying to find out something about human minds and what they contain.

Studying Knowledge of Language

Trying to find out what’s in the mind might seem easy at first. Since knowledge of language is in us—in our minds—shouldn’t we have direct access to it? Shouldn’t we be able to elicit that knowledge by intensive self-reflection—like remembering something forgotten through hard, careful thought? Sorry, things aren’t that simple.

Knowledge of Language Is Tacit

To clarify the problem we face, think about the following sentences, imagining that they are spoken in a natural way, with no word given special emphasis. Concentrate on who is understood as the “surpriser” and the “surprisee” in each:

- (1) Homer expected to surprise him.
- (2) I wonder who Homer expected to surprise him.
- (3) I wonder who Homer expected to surprise.

These sentences are similar in form but curiously different in meaning. Any competent speaker of English will understand sentence (1) to mean that Homer expected to do the surprising and that he expected to surprise someone other than himself. Sentence (2) contains the identical substring of words *Homer expected to surprise him*, but it is immediately understood to have a very different meaning. In fact, it has at least two meanings distinct from that of sentence (1): someone

other than Homer (“who”) is expected to be the surpriser, and the surprisee (“him”) may be either Homer or some third party. Finally, sentence (3) is identical to sentence (2) minus the word *him*, but now Homer again must be the surpriser, rather than the surprisee.

These facts are remarkably intricate and subtle, yet immediately obvious to anyone who has mastered English. But what principles are we following in making these judgments?

In fact, we don’t have a clue—not initially, at least. True, we can make complex judgments about sentences like these. But we cannot directly grasp the basis of our judgments. People don’t consciously know why, when they say *I wonder who Homer expected to surprise him*, the name *Homer* and the pronoun *him* will be taken to refer to different people.

The knowledge that we possess of our language is almost entirely **unconscious** or **tacit knowledge**. In this respect, language appears to be similar to other important parts of our mental life. Sigmund Freud is famous for having proposed that much of the mind’s functioning and contents lies entirely hidden to consciousness. Freud held that unconscious phenomena and processes are no less psychologically real than conscious ones, and that appeal to them is just as necessary for an understanding of human cognition.

I handle unconscious ideas, unconscious trains of thought, and unconscious impulses as though they were no less valid and unimpeachable psychological data than conscious ones. [And] of this I am certain—that anyone who sets out to investigate the same region of phenomena and employs the same method will find himself compelled to take the same position ...
—*Fragment of an Analysis of a Case of Hysteria* (“Dora”), p. 232

Sigmund Freud
1856–1939

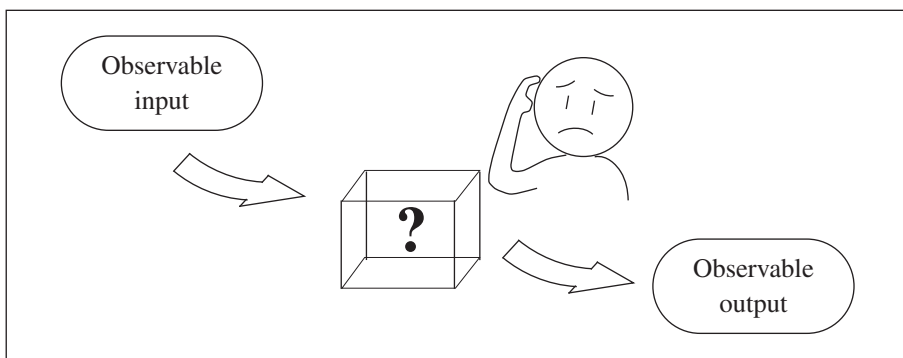


For the most part, the principles and operations behind knowledge of language lie outside the range of consciousness and cannot be recovered by simply sitting down, staring off into space, and thinking hard.

A “Black Box” Problem

If we can’t directly intuit what’s in our minds, then our only option is to approach the investigation of internal things (like knowledge and mental states) as we would approach the investigation of external things (like birds and planets). That is, we must formulate explicit theories about what we know, and we must find ways to test, refine, and extend those theories in order to reach a satisfactory explanation of the facts. Since we can’t look directly at what’s inside the mind, our job will be to figure out what’s inside on the basis of what we can observe from the outside.

Problems of this kind are sometimes called **black box problems**. In a black box problem, we have an unknown mechanism that receives observable input data and produces observable output behaviors. The task is to figure out what’s inside the box on the basis of inputs and outputs alone.



In the case of human language, the observable input is the speech data that people are exposed to as children, the language that they hear around them. The output is their various linguistic behaviors as children and as adults: the sentences and other expressions that they produce, their judgments about their speech and the speech of others, and so on. By carefully examining this kind of information, the linguist must deduce the language mechanism that lies within the human mind.

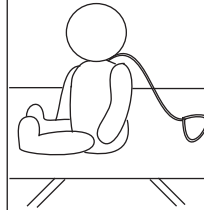
A Talking Analogy

To make the black box nature of the problem more concrete, consider a simple analogy (due to MIT linguist James Harris). For many years, toymakers have produced talking dolls of various kinds. Some have a string on their back or neck that you pull. Others have a button on their wrist or stomach. Still others talk when you talk to them (although these must be turned on initially with a switch).

Imagine yourself an engineer who has been handed a particular model of talking doll: say, the kind that has a string on its neck. Your task (as set by your boss) is to discover exactly how the doll talks. In other words, you have to figure out the properties of the mechanism inside the doll that allows it to do what it does. Suppose also that a certain constraint is placed on your work: you are not allowed to open the doll up and observe the mechanism directly. This makes it a black box problem: you can't look inside.

To solve this problem, you would have to use what's observable from the outside as a basis for guessing what's inside. Examining the doll, you would observe things like this:

- The language mechanism is powered exclusively by pulling the string; there are no plugs or batteries.
- The doll has a fixed repertory of ten or so utterances, which come out in random order (“Mommy, play with me now,” “I want another drink of water,” “I’m sleepy, nite-nite,” etc.).
- All repetitions of a particular utterance are identical.
- The doll always starts at the beginning of an utterance—never in the middle, even if you pull the string out only partway.
- Submerging the doll in water damages the language mechanism.
- The language mechanism is apparently about the size of a tennis ball and is located in the abdominal region.

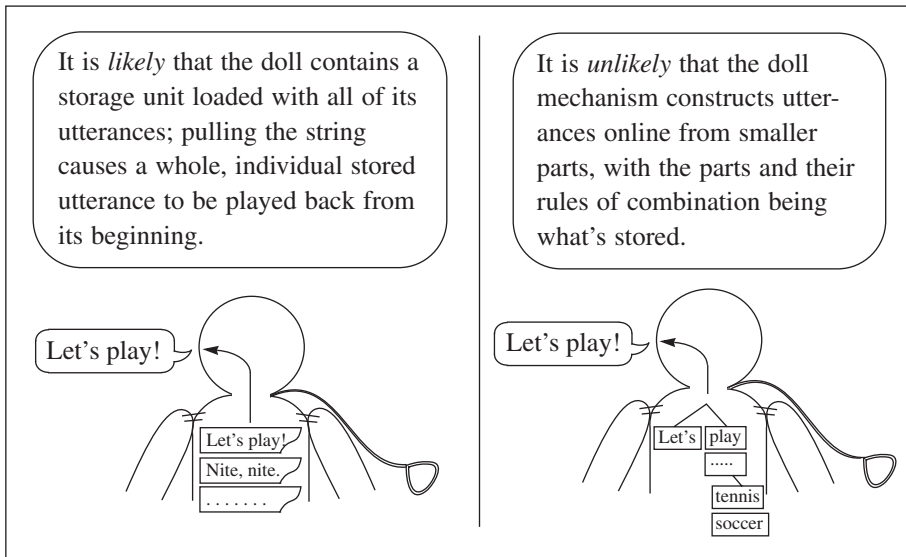


Take a few moments now and write down what mechanism you think is inside the doll, and how these observations imply this mechanism.

Deducing What's Inside the Box from the Output

Thinking about the observable properties of the doll, you can make a pretty good educated guess about what's inside, even if you aren't allowed to cut the doll

open and look inside. For example, since the doll produces a very limited range of utterances and all repetitions of a particular utterance are identical, it is very likely that the utterances are stored within the doll as whole chunks, not constructed online. That is, it is likely that the doll contains a storage unit loaded with all of its utterances; pulling the string causes a whole, individual stored utterance to be played back from its beginning.



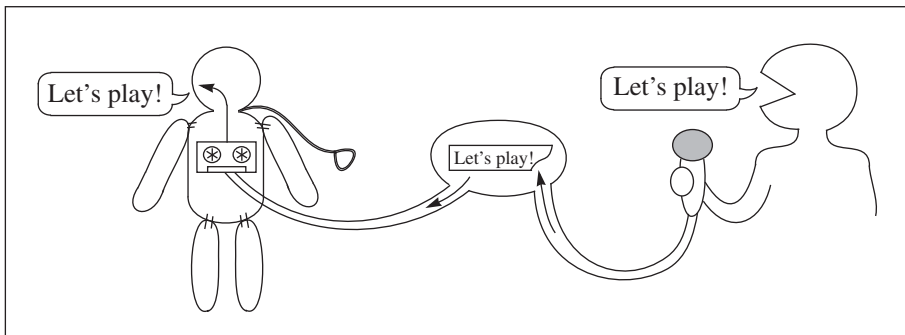
Deducing what's inside humans is vastly more complex than deducing what's inside the doll, but already we can see some things by contrast. For example, since we humans produce an enormous range of utterances, without exact repetitions, it's very unlikely that we have utterances stored within us as whole chunks. Rather, we probably do construct our utterances from smaller parts as we speak, with the parts and their rules of combination being what's stored. With humans, then, something different and more complex is involved. As we will see in later units, the rich complexity of linguistic data—the speech we hear around us, the output we observe—allows us to conjecture a very rich mechanism inside the human mind.

Deducing What's inside the Box from the Input

The data we draw on in solving a black box problem come not only from “output behavior”: in our present case, the utterances produced by talking dolls, or the utterances and linguistic judgments produced by talking humans. They also come

from the input the mechanism receives. Often we can deduce what kind of mechanism is inside the black box by seeing what kind of information initially went into it.

For example, going back to our analogy, suppose you observe that, for the doll, “learning” the ten or so utterances that it produces involves a human being producing each of these utterances. Perhaps you visit the factory where the dolls are made and you observe a person speaking into a microphone that is connected to the doll by a wire. You observe that the doll’s speech exactly repeats that of the person speaking into the microphone, that the utterances the doll ultimately produces are copies of the human’s speech. Such evidence would clearly support your hypothesis that the doll contains some kind of storage and playback device—a disk, a tape player, or something similar. So, the circumstances in which the doll acquires its language can give us information about the mechanism inside it, even when we can’t observe this mechanism directly.



Comparisons with Human Language

Applying this strategy to human language yields surprising results—indeed, some of the most fascinating results in all of the cognitive sciences. Clearly, humans do not learn language like our talking doll, or like a parrot. Although children do repeat expressions that they hear around them in day-to-day speech, often very closely matching the intonation, pitch, and timing of words, their speech goes far beyond what they hear. Children, and indeed humans generally, are extremely creative in their language use, routinely producing utterances they have never encountered before.

Furthermore, the data that form the input to human language acquisition are not clean and precise. Our doll’s utterances were “learned” from very precise, careful speech uttered into a microphone, perhaps in the sheltered environment of a sound booth. But these are not the circumstances in which human speech

is acquired, with careful models of good sentences presented clearly and coherently. In fact, spoken natural language does not provide particularly good models for a child to follow in acquisition. The speech that children hear is often characterized by fragmentary and outright ungrammatical expressions, interruptions, lapses of attention, errors, burps, you name it. When you are listening, speaking, or holding a conversation, your impression is typically one of connected discourse. But that is by no means the reality. The data that children must draw upon in learning a language are remarkably messy and “defective.” (If you need convincing of this, simply lay a tape recorder on a table during a normal daily conversation, and later transcribe three minutes’ worth of the speech you have recorded. How many complete, coherent, and grammatical sentences do you observe?)

Finally, the evidence that children draw upon in learning language is at best extremely indirect. Recall our three example sentences (repeated here):

- (1) Homer expected to surprise him.
- (2) I wonder who Homer expected to surprise him.
- (3) I wonder who Homer expected to surprise.



The judgments we make about “surpriser” and “surprisee” are intricate and subtle, but obvious to anyone who knows English.

How did we learn the principles that underlie these judgments? Surely they were not taught to us directly or explicitly. They are not found in any English grammar textbook; they have never even been noticed, except by a minuscule circle of specialists, and indeed, they are still not known with absolute certainty even by specialists. Yet every normally developing English-speaking child masters them at an early age with no special effort.

Universal Grammar

From these reflections, it is clear that language learning and its outcome present a surprising picture. Our resulting knowledge of language has these properties:

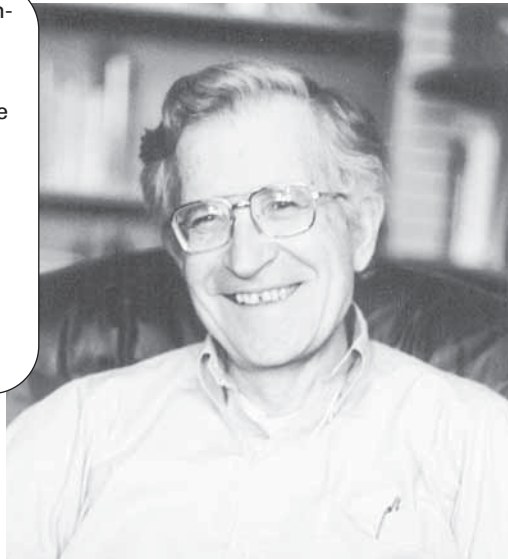
- It is **tacit**; we come to know many things that we don't know that we know.
- It is **complex**; it underwrites very subtle and intricate judgments.
- It is **untutored**; the vast bulk of it was never taught to us directly.
- It is **gained in the face of very impoverished input**.

One plausible explanation for this picture—perhaps the only plausible explanation—has been proposed by the linguist Noam Chomsky. Chomsky suggests that children come to the task of language acquisition with a rich conceptual apparatus already in place that makes it possible for them to draw correct and far-reaching conclusions on the basis of very little evidence. Human language learning involves a very powerful cognitive system that allows learners to infer their grammar from the meager data they are presented with in day-to-day speech. Chomsky terms this cognitive system **Universal Grammar**, or **UG** for short.

We may think of Universal Grammar as the system of principles that characterizes the class of possible grammars by specifying how particular grammars are organized (what are the components and their relations), how the different rules of these components are constructed, how they interact, and so on. ... Universal Grammar is not a grammar, but rather ... a kind of schematism for grammar.
—*Language and Responsibility*, pp. 180, 183

Noam Chomsky
Institute Professor
Massachusetts Institute of
Technology

Photo by Donna Coveney/MIT.
Reprinted with permission.



UG in humans is very roughly analogous to the mechanism inside our talking doll. Although the doll's device is not a deductive conceptual mechanism, it is one that allows dolls equipped with it to “learn” or at least be made to “speak” any language. By simply recording utterances in one or another language on the

disk, drum, tape, or whatever device the mechanism uses for storing its messages, dolls can be made to utter sentences of German, Hindi, Maori, and so on. Furthermore, just as the doll's mechanism is part of its basic physical structure, is specific to that kind of doll, and is found in all dolls of that kind, so too the basic mechanism that makes it possible for humans to learn language is apparently part of our physical structure (our genetic endowment), is peculiar to the human species alone, and is found in all members of our species (putting aside cases of pathology).

Evidence for Universal Grammar

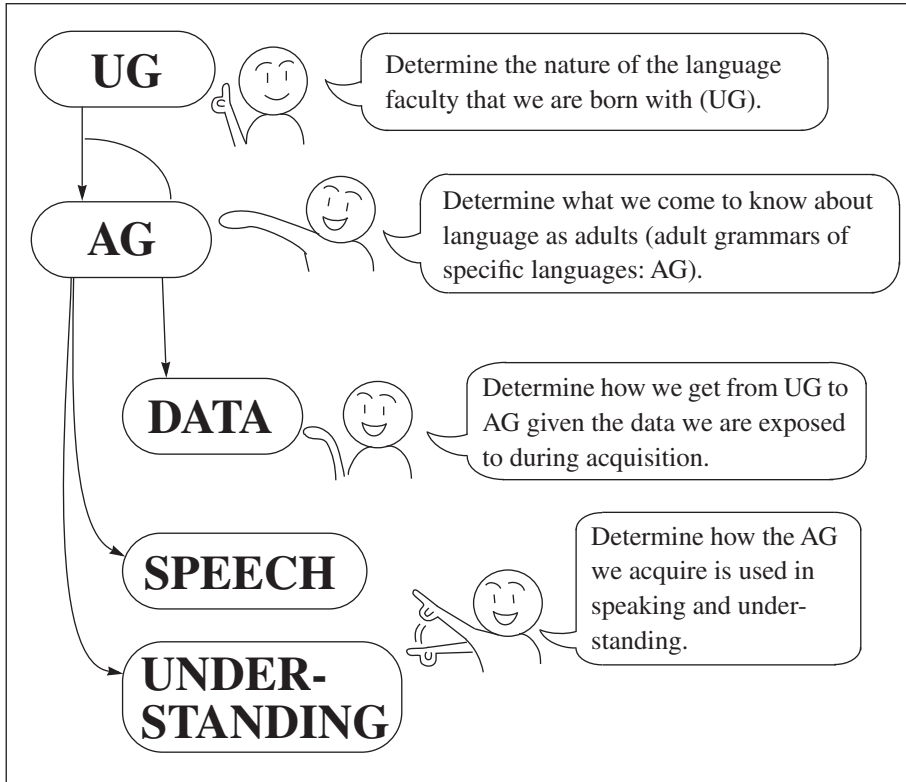
Evidence for basic linguistic endowment in humans comes from at least three sources:

- The acquisition process is surprisingly uniform for all children, even though the languages being learned may seem wildly different.
- Although the languages acquired by children are superficially diverse, deeper investigation reveals significant, shared design features.
- With equal facility and with no special training, all children, of whatever ethnic or genetic background, learn whatever language or languages they have significant contact with. No one has a racial or genetic predisposition to learn one language more readily than another.

These facts would be all but impossible to understand if normally developing human children did not come to the task of native-language acquisition equipped with a single standard acquisition device, provided by their biological makeup.

The Task of Linguistics

Given these results, we can reformulate the task of linguistics in investigating knowledge of language. Linguistics must accomplish the following:



Review

1. Linguistics addresses knowledge of language. It seeks to answer three basic questions.



- What exactly do we know when we know a language?
- How do we acquire that knowledge?
- How do we use that knowledge?

2. We figure out what's in people's minds by deducing it from the data they are exposed to and the behavior they exhibit.

It's a black box problem!



3. We know many complicated things about our language that we were never directly taught. Moreover, the data from which we draw our knowledge are often defective.

This suggests that some sort of mechanism must already be in place that supports language acquisition.



4. Part of language we know as children, prior to experience. It is with us at birth, as part of our genetic endowment as human beings.

It's called **Universal Grammar (UG)**!



Dividing Up the Problem Area

In studying people's knowledge of language, modern linguistics follows a general methodological principle set down by the French philosopher René Descartes. Descartes counseled that in approaching any problem, we should begin by trying to divide it up into smaller, more manageable parts.

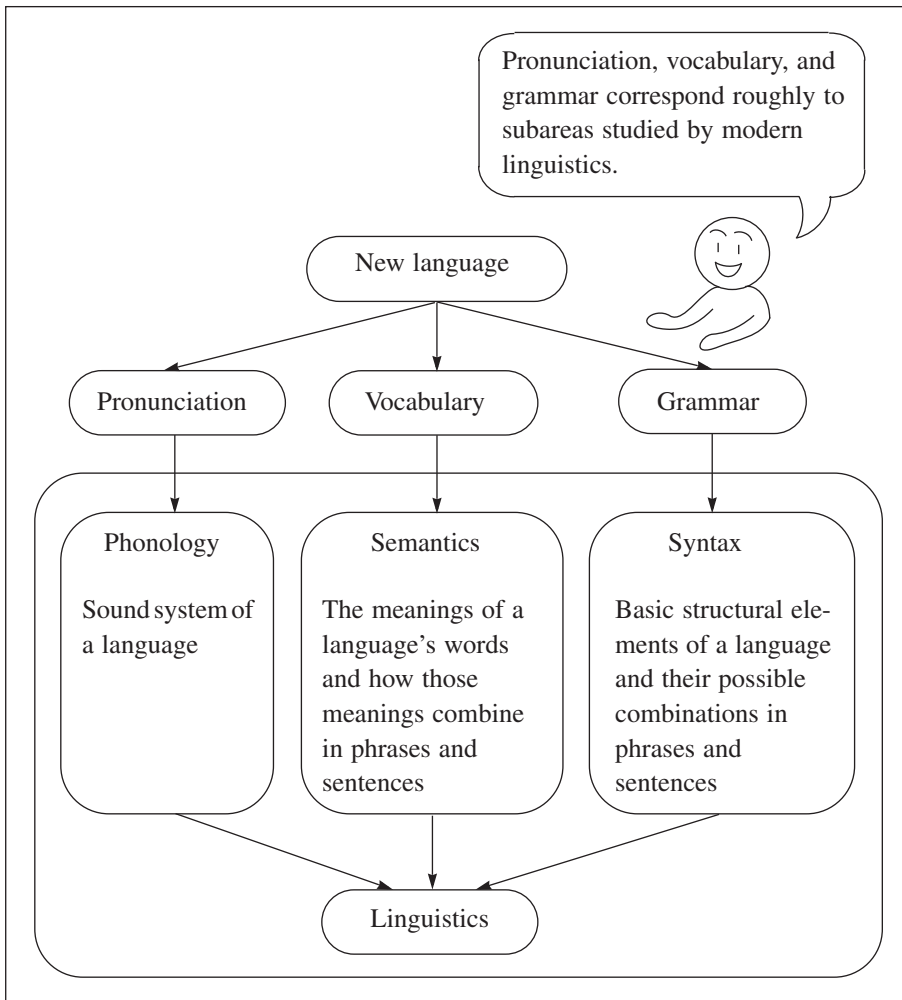
We should divide a problem into as many parts as admit of separate solution.

—*Discourse on Method*, p. 92

René Descartes
1596–1650



When you study a new language, there are a number of things you must master, including pronunciation, vocabulary, and grammar. These can be viewed as separate parts of your developing linguistic knowledge, and they correspond approximately to the parts of linguistic knowledge studied by the modern field of linguistics:



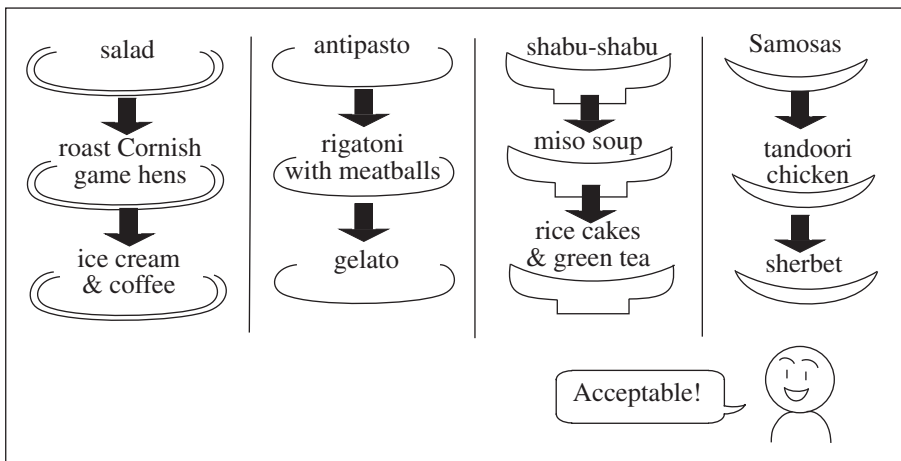
Syntax in particular studies and describes what people know about the *form* of the expressions in their language. It studies the basic grammatical patterns of language and what gives rise to them.

How do we go about describing what people know about grammatical patterns? To gain some insight into this, let's start with the broader question of how we capture patterns in any domain. We'll pursue it in relation to a question that's always close to our hearts (and stomachs): what's for dinner?

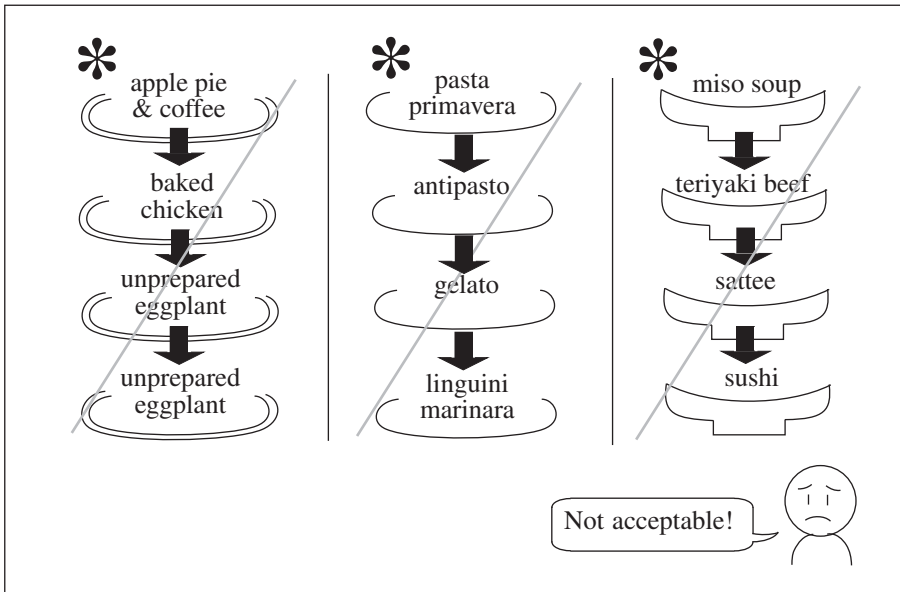
Capturing Patterns: What's for Dinner?

We're all used to eating meals on the fly these days: a quick sandwich and soda at a deli, or perhaps a fresh salad from a salad bar if we're eating healthy. In casual meals of this kind, there are few constraints on what can be eaten or the order in which it's consumed. Pretty much anything goes. However, when it comes to a real “sit-down meal”—the sort of thing you might invite a friend over to your house for—most people have definite feelings about what constitutes a proper dinner: what it can and should include, and what form it should take.

For example, depending on your nationality or cultural heritage, here are some possible meals that you might feel to be acceptable:



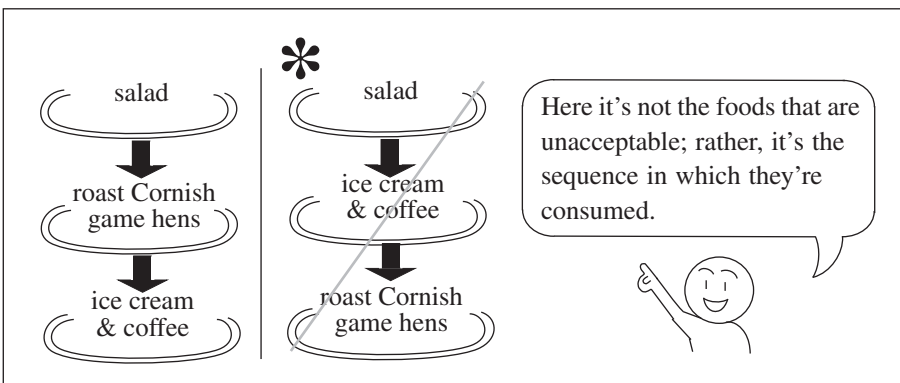
By contrast, most people would reject menus like these (marked with an asterisk “*” —sometimes called a “star” in linguistics—to indicate that they are unacceptable):



What? In What Order? In What Combinations?

Some of our intuitions about what makes an acceptable meal concern *what* we eat. For example, traditional American meals don't include unprepared vegetables of certain kinds like eggplant or parsnips. Nor do they include raw fish—fish that isn't cooked, smoked, or salted in some way.

Other intuitions about what makes an acceptable meal concern the *order* in which we eat various dishes. For example, whereas the first menu here is acceptable, the second isn't:

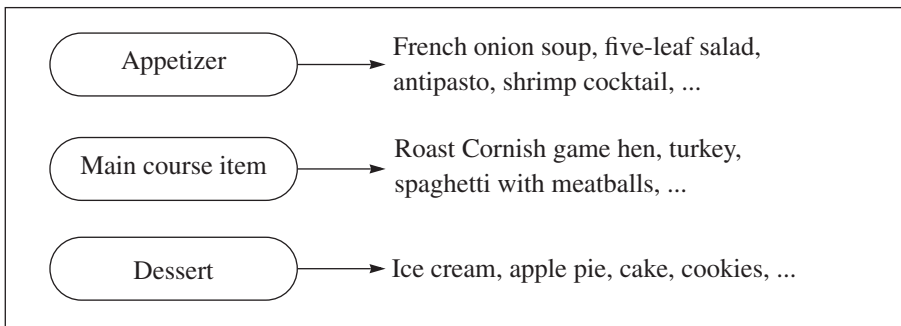


Finally, there are constraints on what *combinations* of things should appear together in a single meal. For example, while sattee, sushi, and teriyaki beef are all fine items on a Japanese menu, in Japanese culture they probably wouldn't be eaten all together in a single meal. In the same way, in American culture baked chicken and hot dogs wouldn't be eaten together—a meal would include one or the other, but not both.

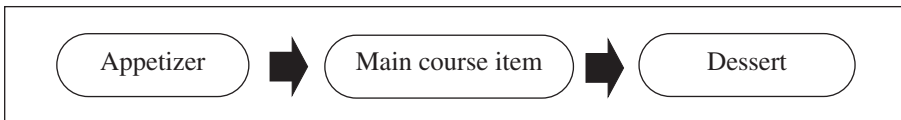
Categories and Arrangements

Suppose you were asked to describe what constitutes an acceptable or “well-formed” traditional American meal—that is, to work out the pattern behind possible American dinners. How would you go about it?

One natural idea would be to divide the various foods into categories and subcategories. If you look at the suggested menus in a traditional cookbook, you will find terms like *appetizer*, *main course item*, and *dessert* (categories). The various foods (subcategories) can be classified according to these categories:



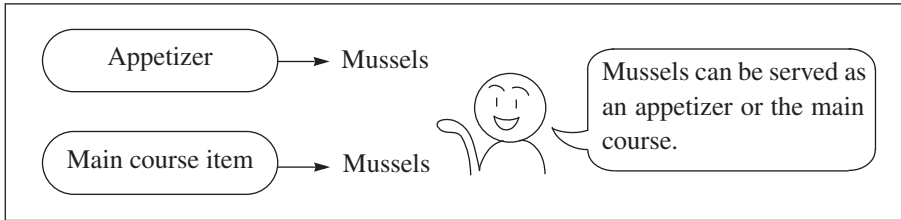
With this classification, you could then state the pattern of an acceptable American meal in terms of the arrangements of these general categories. For example, you might say that a possible dinner has the following general pattern:



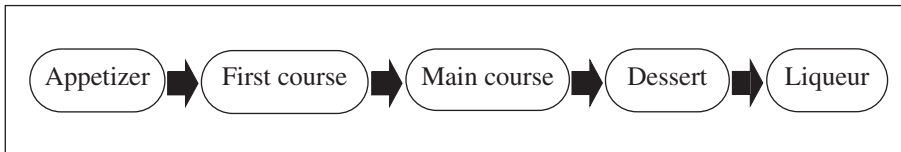
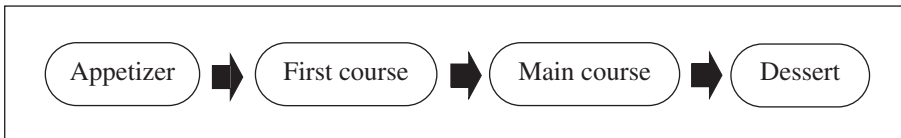
This strategy would capture what is eaten (the things in the categories), the order in which they are eaten (expressed by the order of the general categories), and the combinations.

Of course, many subtleties could come into play at this point. For example, some foods can occur in more than one category. Many main course items like

shellfish can also be served as appetizers as long as the portion is small enough. You might want to classify such foods as both appetizers and main course items:



A very formal meal might include a first course or a fish course before the main course and possibly liqueur after dessert. This means that you would have to add items to the general pattern:



To summarize: there are numerous factors to consider in describing an American meal completely. Foods have to be cross-classified to some extent, and there is a (potentially large) number of patterns to account for. Nonetheless, the basic procedure used here appears sound and capable of being extended to these other cases without too much difficulty.

Capturing Syntactic Patterns

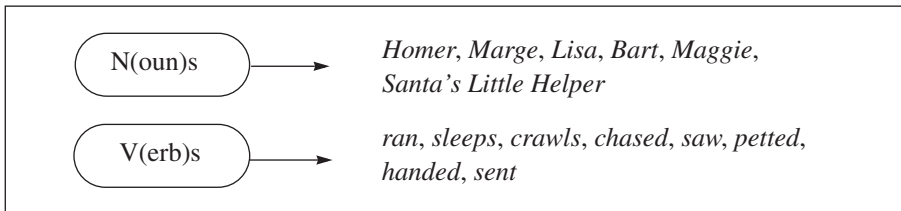
The example of eating patterns suggests a general strategy for capturing all patterns that hold for some collection of objects. We proceed as follows:

- Classify the objects into general categories.
- State the possible patterns that we observe as arrangements of the general categories.

Let's try applying this lesson to sentence patterns using the following simple grammatical data. The three lists contain both acceptable and unacceptable sentences; the unacceptable ones are marked with an asterisk.

I	II	III
Bart ran.	Homer chased Bart.	Homer handed Lisa Maggie.
Homer sleeps.	Bart saw Maggie.	Marge sent Bart SLH.
Maggie crawls.	Maggie petted SLH.	*Sent Marge Bart SLH.
*Ran Maggie.	*Chased Bart Homer.	*Marge Bart SLH sent.
*Crawls Homer.		

Following the strategy suggested above, we might begin by classifying the expressions in I–III into different general categories. Just as traditional cookbooks separate foods into different menu items like appetizer and main course, traditional grammar books separate the words into different **parts of speech**. Parts of speech represent general categories of words. Traditional parts of speech include categories like noun, verb, preposition, adjective, and article. For present purposes, the two traditional categories of noun and verb will suffice for dividing up all the words in I–III:



Next, just as we analyzed acceptable patterns of meals into sequences of general categories of foods, we analyze the acceptable patterns of English sentences into sequences of our general categories of words:

Acceptable English sentences (I):	N V
Acceptable English sentences (II):	N V N
Acceptable English sentences (III):	N V N N

As in the case of meals, these rules state what can appear (the words in the categories), the order in which they appear (expressed by the order of the general

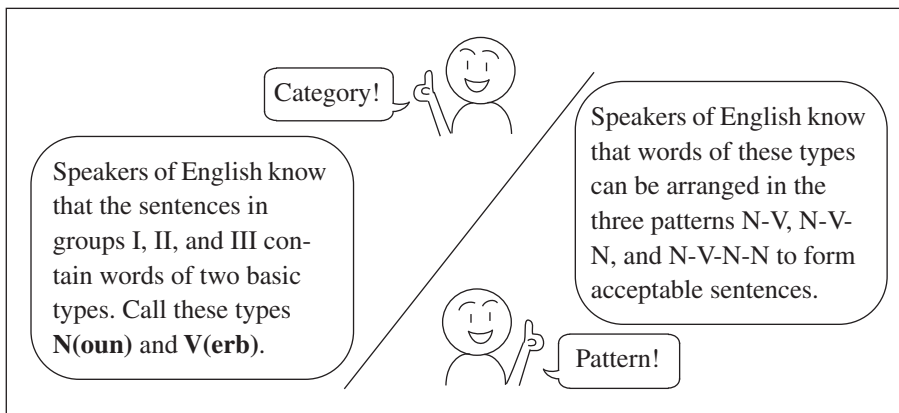
categories), and their possible combinations (expressed by what's in the separate categories).

Once again, there are many additional points and subtleties. Just like some foods, certain words seem to occur in more than one category. For example, the sequence of sounds that we pronounce “saw” can appear as a noun, as in *The saw was old*, or as a verb, as in *Bart saw Maggie*. Furthermore, just as there are additional menu items and patterns beyond appetizer–main course–dessert, there are many additional categories of words (adverbs, intensifiers, conjunctions, determiners, etc.) and many patterns of categories beyond those just considered.

These don't seem to raise any problems of principle, however. As before, the basic procedure appears sound and capable of being extended to other cases. We simply introduce new words, new categories, and new patterns.

Speakers Know Patterns

The results above allow us to formulate an explicit hypothesis about what speakers know when they have systematic knowledge of some structured domain. We could hypothesize that they know **categories** and **patterns**. In the case of sentence patterns, we would be making the following conjecture:

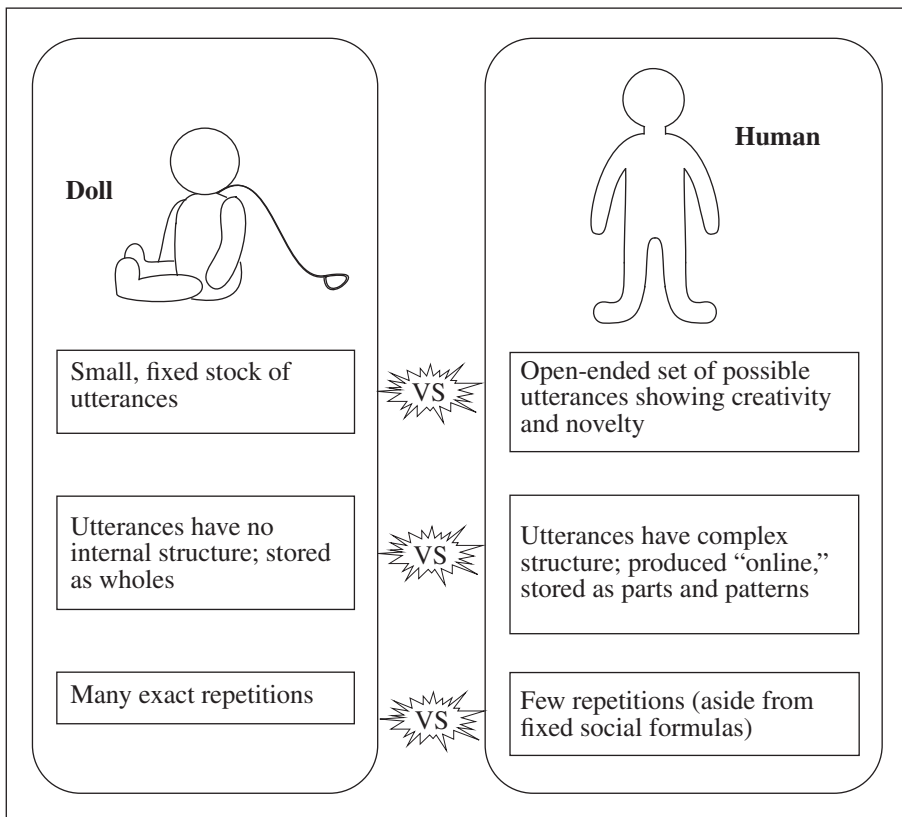


This would be the kind of knowledge that a syntactician might reasonably attribute to speakers of English. Attributing this type of knowledge to speakers constitutes an explicit proposal about (part of) what those speakers know about the structure of their language.

Internal Structure

The hypothesis that speakers know categories and patterns entails that their knowledge of syntax is structured in a certain way. Our explanation for how English speakers are able to recognize well-formed sentences involves seeing those sentences as divided into parts that are arranged in certain definite ways. The hypothesis states that a well-formed sentence of English is composed of nouns and verbs, and it is the way these parts are arranged that determines well-formedness.

There is strong evidence that our grasp of syntax must be like this: structured out of parts. To appreciate this, recall the properties distinguishing a human’s linguistic behavior from that of a talking doll:



As we saw, a talking doll produces a small number of utterances, usually no more than ten or twelve; and each repetition of a given utterance is identical to any

other (ignoring wear and tear on the doll). On this basis, we quickly concluded that the doll's linguistic mechanism must be some form of playback device, in which each utterance the doll can produce is stored as a separate unit.

Human linguistic capacities are nothing like this, however. For one thing, human linguistic competence allows us (at least in principle) to produce infinite collections of well-formed sentences. Consider, for example, this set of sentences (from Platts 1979, p. 47):

The horse behind Pegasus is bald.
 The horse behind the horse behind Pegasus is bald.
 The horse behind the horse behind the horse behind Pegasus is bald.
 The horse behind the horse behind the horse behind the horse behind Pegasus is bald.
 ...




Clearly, this list could be extended indefinitely—it has infinitely many members.

Although this set of sentences is infinite, English speakers recognize that every sentence in the set is a well-formed sentence of English. Of course, our actual capacity to produce or process sentences like these is limited in certain ways. When the sentences get too long, we can't get our minds around them: we forget how they began, or we get distracted, or we simply lose track. Consequently, we can't show our mastery of them in the usual ways. But it seems that these limitations reflect constraints on such things as memory and attention span and have little to do with specifically linguistic abilities. If we had unlimited attention spans, life spans, memories, and so on, we would presumably be able to produce all the sentences in the set.

The infinite size of such collections shows that unlike the doll's mechanism, our minds don't simply store the sentences that we produce and understand as separate units. Our brains are finite objects with finite storage capacity. One simply cannot get an infinite object into a finite brain. On the other hand, if sentences are structured, and built up out of smaller parts, then our ability to produce an infinite number of sentences can be explained. Suppose we know a basic stock of words and a basic stock of patterns for combining them. Suppose further that we are able to reuse patterns in the process of constructing of a sentence. Then this will be enough to produce an infinite set:

The horse behind Art N P	Pegasus is bald. N		
The horse behind Art N P	the horse behind Art N P	Pegasus is bald. N	
The horse behind Art N P	the horse behind Art N P	the horse behind Art N P	Pegasus is bald. N

Notice that our infinite collection of Pegasus sentences involves reusing the Art-N-P pattern!



By drawing on this pattern over and over again, we are able to construct sentences of greater and greater length—indeed, of potentially any length. Again, all of this points to the central importance of categories and patterns—parts and structures—in a sentence.

EXERCISES

1. Give four important properties that distinguish human linguistic abilities from those of a talking doll like Chatty Cathy® or Teddy Ruxpin®.
2. Human knowledge of language shows four key properties. What are they?
3. What is our strategy for capturing the syntactic patterns that hold across the sentences of a language?
4. State the categories found in sentences (1)–(4) and the pattern(s) of combining these categories:
 - (1) Homer came home tired.
 - (2) Homer heard Maggie clearly.
 - (3) Lisa picked Maggie up.
 - (4) Marge thinks Bart chased Lisa.
5. The following set of sentences is potentially infinite, making use of a recurring pattern. What is the pattern?
 - (1) Bart laughed.
Bart laughed and-then Bart laughed again.
Bart laughed and-then Bart laughed again and-then Bart laughed again.
Bart laughed and-then Bart laughed again and-then Bart laughed again and-then Bart laughed again.
...
6. The following examples are from Japanese. Assume that the Japanese parts of speech are the same as the parts of speech of the English gloss. What is the pattern? (Note: The little particles *-ga*, *-o*, and *-ni* are used in Japanese to indicate a word's status as a subject, direct object, or indirect object, respectively.)
 - (1) Taroo-ga Pochi-o mita.
Taroo-NOM Pochi-ACC saw
'Taroo saw Pochi.'
 - (2) Taroo-ga Hanako-ni Pochi-o ageta.
Taroo-NOM Hanako-DAT Pochi-ACC gave
'Taroo gave Pochi to Hanako.'

PART II Grammars as Theories

It's time for our first “map check”—a stop to consider where we are in the larger landscape, and what to look for in the landscape ahead.

The urge toward science typically starts with **phenomena that raise questions**. We observe something that surprises us and makes us curious. We want to know more. We start asking questions. The phenomena that surprise us needn't be exotic or technical—things found only in laboratories or observed with complex apparatus. The everyday world presents us with many puzzles.


It is important to learn to be surprised by simple things. ... The beginning of a science is the recognition that the simplest phenomena of ordinary life raise quite serious problems: Why are they as they are, instead of some different way?
—*Language and Problems of Knowledge*, p. 43




Human language is like this. Language is something that surrounds us and that we take for granted in daily life. But as we have seen, when we reflect carefully on our knowledge of language and pose even the most basic questions about it, we become surprised and puzzled!

- What do we know when we know a language?
- How did we come to know it?
- How do we use that knowledge?

A certain intellectual effort is required to see how such phenomena can pose serious problems or call for intricate explanatory theories. One is inclined to take them for granted as necessary or somehow “natural.”
—*Language and Mind*, p. 24



Surprises, puzzles, and questions unsettle us. They capture our attention and occupy our thoughts. They press us to **construct a theory** (or story) about what is going on—one that will solve the puzzles, answer the questions, and put our minds at rest. Science does not end with theory construction, however. A hallmark of science is its drive to **test theory against experience**.



A scientist, whether theorist or experimenter, puts forward statements, or systems of statements, and tests them step by step. In the field of the empirical sciences ... [the scientist] constructs hypotheses, or systems of theories, and tests them against experience by observation and experiment.
—*The Logic of Scientific Discovery*, p. 27

Sir Karl Popper
1904–1994

Theories that survive repeated testing (what Popper called the “clash with reality”) are theories in which we gain increasing confidence.

These points chart the general path ahead for us. We have identified some puzzling and intriguing questions about our knowledge of language. Our task now

is to begin constructing a theory that will address these questions and illuminate the phenomena that raise them. Furthermore, we must find ways of testing our theory against experience, to see whether it's correct. Indeed, we have already begun this process. Our initial observations of human language have already ruled out a theory in which it consists of a store of complete sentences, like the talking doll's "language."

A natural starting point is one of the questions raised earlier: exactly what do we know when we know the syntax of our language? To anticipate slightly, this part of the book will develop the idea that **people know a grammar**, conceived of as a set of rules and principles. In this view, **a grammar constitutes a scientific theory about (a part of) human linguistic knowledge**. The general questions confronting us will therefore include these:

- How do we systematically construct a grammar?
- How do we test it?
- How and when do we revise and extend it, in response to our tests?

To aid their investigations, scientific researchers often construct tools (physical or conceptual) to make inquiry easier, more efficient, or more precise. In the next unit, we will look at some basic tools that will assist us in grammar building.

Introducing Phrase Structure Rules

Review

1. Syntax studies speakers' knowledge of the structural arrangement of words and phrases in their language.

The pattern of its forms!



2. We capture patterns using categories and their arrangements.



Divide words into categories and state sentence patterns in terms of them!

3. Speakers' knowledge of linguistic patterns must be structured like this.



It must define a well-formed sentence in terms of the form and arrangements of smaller constituent bits.

Generating Sentences

So far we've described syntactic patterns by writing out statements like "N V N is an acceptable pattern for a sentence of English." Let's now start using some simple notation for this purpose. We will adopt the arrow notation on the left-hand side below as a shorthand way of saying what is written out on the right-hand side:

Notation	English prose
$N \rightarrow \textit{Homer}$	“ <i>Homer</i> is a noun.”
$N \rightarrow \textit{Marge}$	“ <i>Marge</i> is a noun.”
$N \rightarrow \textit{Lisa}$	“ <i>Lisa</i> is a noun.”
$N \rightarrow \textit{Bart}$	“ <i>Bart</i> is a noun.”
$N \rightarrow \textit{Maggie}$	“ <i>Maggie</i> is a noun.”
$N \rightarrow \textit{Santa's Little Helper}$	“ <i>Santa's Little Helper</i> is a noun.”
$V \rightarrow \textit{ran}$	“ <i>Ran</i> is a verb.”
$V \rightarrow \textit{sleeps}$	“ <i>Sleeps</i> is a verb.”
$V \rightarrow \textit{crawls}$	“ <i>Crawls</i> is a verb.”
$V \rightarrow \textit{chased}$	“ <i>Chased</i> is a verb.”
$V \rightarrow \textit{saw}$	“ <i>Saw</i> is a verb.”
$V \rightarrow \textit{petted}$	“ <i>Petted</i> is a verb.”
$V \rightarrow \textit{handed}$	“ <i>Handed</i> is a verb.”
$V \rightarrow \textit{sent}$	“ <i>Sent</i> is a verb.”
$S \rightarrow N V$	“A noun followed by a verb is a sentence (of English).”
$S \rightarrow N V N$	“A noun followed by a verb followed by a noun is a sentence (of English).”
$S \rightarrow N V N N$	“A noun followed by a verb followed by a noun followed by another noun is a sentence (of English).”

One virtue of this arrow notation is brevity. It's a lot quicker to write “ $S \rightarrow N V N N$ ” than it is to write out “A noun followed by a verb followed by a noun followed by another noun is a sentence (of English).”


Patterns as Rules

Another virtue of the arrow notation is that it suggests a kind of “recipe” or procedure for constructing English sentences. That is, we can view the statements above as rules that can be followed to construct well-formed English clauses.


Consider the following procedure:

1. Write down the symbol "S". Interpret a statement " $X \rightarrow Y Z$ " as an instruction to replace or rewrite the symbol X with the symbol Y followed by the symbol Z.
2. Whenever you have two or more rules for rewriting the same symbol, choose freely among them.

With this procedure, we can use our rules to produce a large number of well-formed English sentences by a series of rewritings:

 **EXAMPLE** Maggie crawls.

Start	<input type="text" value="S"/>		write down the symbol "S"
Step 1	<input type="text" value="N"/>	<input type="text" value="V"/>	rewrite "S" using " $S \rightarrow N V$ "
Step 2	<input type="text" value="Maggie"/>	<input type="text" value="V"/>	rewrite "N" using " $N \rightarrow \textit{Maggie}$ "
Step 3	<input type="text" value="Maggie"/>	<input type="text" value="crawls"/>	rewrite "V" using " $V \rightarrow \textit{crawls}$ "

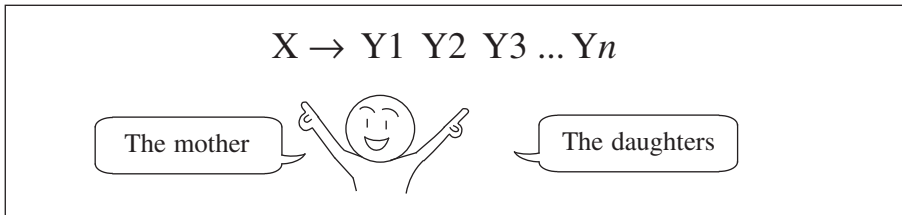
 **EXAMPLE** Homer chased Bart.

Start	<input type="text" value="S"/>		write down the symbol "S"	
Step 1	<input type="text" value="N"/>	<input type="text" value="V"/>	<input type="text" value="N"/>	rewrite "S" using " $S \rightarrow N V N$ "
Step 2	<input type="text" value="Homer"/>	<input type="text" value="V"/>	<input type="text" value="N"/>	rewrite "N" using " $N \rightarrow \textit{Homer}$ "
Step 3	<input type="text" value="Homer"/>	<input type="text" value="chased"/>	<input type="text" value="N"/>	rewrite "V" using " $V \rightarrow \textit{chased}$ "
Step 4	<input type="text" value="Homer"/>	<input type="text" value="chased"/>	<input type="text" value="Bart"/>	rewrite "N" using " $N \rightarrow \textit{Bart}$ "

The end product in each case is a well-formed English sentence. The rules furnish a procedure for generating English sentences: a **generative procedure**.

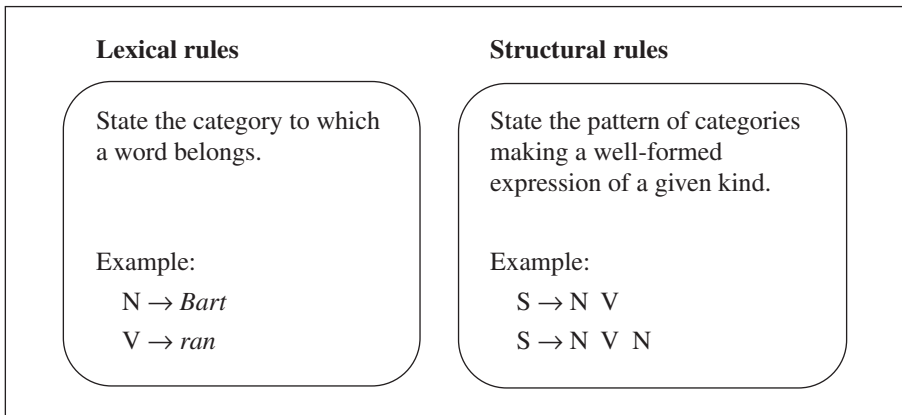
Phrase Structure Rules

Rules of the kind given above are called (**context-free**) **phrase structure rules** (or **PS rules** for short). They have the general form shown below:



This says that the single symbol X can be rewritten as the string of symbols $Y_1 Y_2 Y_3 \dots Y_n$. Since the symbol X is understood as giving rise to the symbols $Y_1 Y_2 Y_3 \dots Y_n$, the former is sometimes spoken of as the **mother** of the latter; alternatively, the latter are spoken of as the **daughters** of the former.

The phrase structure rules listed above can be divided into two basic kinds:



Tree Diagrams and Derivations

We have seen that the generative procedure yields a derivation like this for *Maggie crawls*:

Start	S		write down the symbol “S”
Step 1	N	V	rewrite “S” using “S → N V”
Step 2	Maggie	V	rewrite “N” using “N → <i>Maggie</i> ”
Step 3	Maggie	crawls	rewrite “V” using “V → <i>crawls</i> ”

Notice that in deriving this sentence we could have applied our rules in a different order:

Start	S		write down the symbol “S”
Step 1	N	V	rewrite “S” using “S → N V”
Step 2	N	crawls	rewrite “V” using “V → <i>crawls</i> ”
Step 3	Maggie	crawls	rewrite “N” using “N → <i>Maggie</i> ”

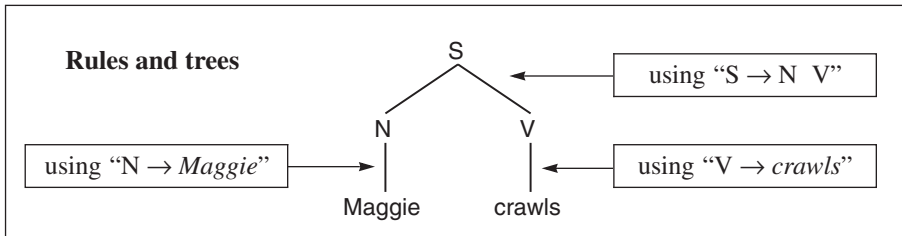
If you think about it, you’ll see that any set of rules will produce a **family of derivations** that differ by applying the rules in different orders. Thus, *Maggie crawls* has two different derivations under our rules. *Homer chased Bart* has nine different derivations. And so on.

Generating Tree Diagrams

There is a useful way of abbreviating the derivations for a sentence produced under a set of rules: with a **phrase marker** or **tree diagram**. Suppose we do this:

1. Write down the symbol S.
2. Pick any rule that can be used to rewrite S (any rule of the form “ $S \rightarrow \dots$ ”).
3. Write the symbols that appear on the right-hand side of the rule beneath S and connect them to S by lines.
4. Repeat the procedure with the symbols that now appear beneath S (that is, pick a rule that can be used to rewrite them; write the symbols occurring on the right-hand side of their rules beneath them and connect with lines).
5. Continue this way until no more symbols can be added.

The result is a tree diagram with S at the top, with branches in the middle, and with words at the bottom.



The string of words at the bottom of the tree is the sentence that we are trying to generate. It is sometimes called the **terminal string** of the tree.

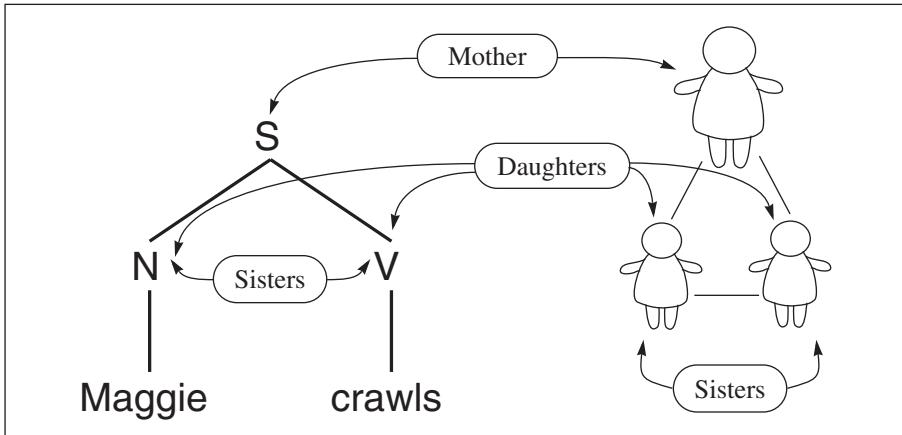
Tree diagrams display how a grammar generates a given sentence like *Maggie crawls*, ignoring the order in which the rules are applied. A tree diagram therefore abbreviates the family of alternative derivations that differ only in order of rule application.

Some Terminology

We will be using tree diagrams a great deal in the units that follow, so it is useful to have some terminology for talking about them. The points in a tree that are labeled by categories like S, N, and V or words like *Homer* and *Lisa* are called the **nodes** of the tree. The lines that connect nodes are called the **branches** of the tree. The single node at the top is called the **root node** of the tree. And the nodes at the very ends of the branches—the words—are called the **terminal nodes** or **leaf nodes** of the tree.

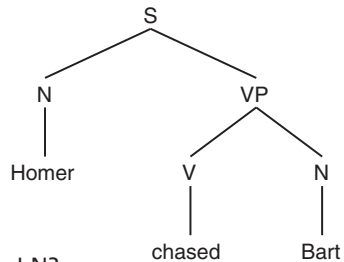
We also need terminology for talking about a given node in relation to other nodes in a tree. The node that appears immediately above a given node is its **mother node**. The nodes that appear immediately below a given node are its

daughter nodes. As in real genealogical trees, a node can have at most one mother, but it can have more than one daughter. Two nodes that have the same mother—two daughters of the same mother—are called **sister nodes**:



 EXERCISE

Answer the following questions for the tree at the right:



1. What is the root node?
2. What are the leaf nodes?
3. Which node is the mother of *Homer*?
4. Which node is the daughter of *Homer*?
5. Which node is the mother of the right-hand N?
6. Which node is the daughter of the right-hand N?
7. Which node is the sister of the left-hand N?
8. Which node is the sister of *chased*?
9. Which nodes are the daughters of S?

Tree Diagrams and Rules

There is a close correspondence between trees and the rules that are used to produce them. If you are given a set of rules and a tree, it's easy to determine

whether the rules generate the tree. Likewise, if you are given a tree, you can easily determine a set of rules that would generate it.

? QUESTION Do the rules in (i) generate the tree in (ii)?

(i)

$$S \rightarrow N VP$$

$$VP \rightarrow V N$$

$$N \rightarrow \textit{Homer}$$

$$V \rightarrow \textit{chased}$$

$$N \rightarrow \textit{Bart}$$

(ii)

```

graph TD
    S --> N1[N]
    S --> VP[VP]
    N1 --> Homer[Homer]
    VP --> V[V]
    VP --> N2[N]
    V --> chased[chased]
    N2 --> Bart[Bart]
  
```

ANSWER Yes. To verify this, we check each node and its daughters, to see that there is a corresponding rule.

Step 1 Start with the top node, S. It has the two daughters N and VP. To produce this part of the tree, we therefore need $S \rightarrow N VP$ in our set of rules. **There is such a rule!**

Step 2 Go on to the N node. It has the single daughter *Homer*. To produce this part of the tree, we need a rule $N \rightarrow \textit{Homer}$. **There is such a rule!**

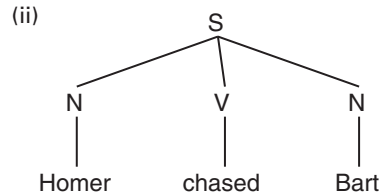
Step 3 Next take the VP node. It has the two daughters V and N. To produce this part of the tree, we need a rule $VP \rightarrow V N$. **There is such a rule!**

Applying this reasoning to the remaining two nodes, you'll see that they check out too. Since every mother-daughter part of the tree corresponds to a rule in the list, the tree can be generated by the list.

? QUESTION

Do the rules in (i) generate the tree in (ii)?

- (i) $S \rightarrow N VP$
 $VP \rightarrow V N$
 $N \rightarrow Homer$
 $V \rightarrow chased$
 $N \rightarrow Bart$



ANSWER

No! Which mother-daughter parts of the tree fail to correspond to rules in the list?

Syntactic Ambiguity

We've seen that sentences can have more than one derivation under a given set of rules if we simply apply the rules in different orders. This is not the only way for multiple derivations to arise, however. Consider the following set of rules and the sentence beside it:

Rules

- $S \rightarrow N V N$
 $S \rightarrow N VP$
 $VP \rightarrow V N$
 $N \rightarrow Homer$
 $V \rightarrow chased$
 $N \rightarrow Bart$

Sentence

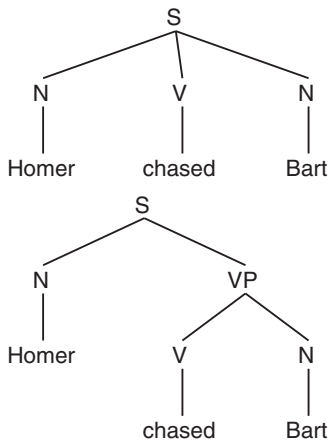
Homer chased Bart.

? QUESTION

How many different tree diagrams can you write for this sentence given these rules?

ANSWER

The rules allow *two* different tree diagrams:



Tree diagrams collapse together derivations that use the *same* rules in different orders. But under this grammar, *Homer chased Bart* has derivations that use *different* rules!



There is a (family of) derivation(s) in which we use the rule $S \rightarrow N V N$ to rewrite the S symbol:

Start	S			write down the symbol "S"
Step 1	N	V	N	rewrite "S" using " $S \rightarrow N V N$ "
Step 2	Homer	V	N	rewrite "N" using " $N \rightarrow Homer$ "
Step 3	Homer	chased	N	rewrite "V" using " $V \rightarrow chased$ "
Step 4	Homer	chased	Bart	rewrite "N" using " $N \rightarrow Bart$ "

But there is also another (family of) derivation(s) in which we use the rules $S \rightarrow N VP$ and $VP \rightarrow V N$:

Start	S				write down the symbol “S”
Step 1	N	VP			rewrite “S” using “ $S \rightarrow N VP$ ”
Step 2	Homer	VP			rewrite “N” using “ $N \rightarrow Homer$ ”
Step 3	Homer	V	N		rewrite “VP” using “ $VP \rightarrow V N$ ”
Step 4	Homer	chased	N		rewrite “V” using “ $V \rightarrow chased$ ”
Step 5	Homer	chased	Bart		rewrite “N” using “ $N \rightarrow Bart$ ”

These two different (families of) derivations correspond to two different tree diagrams because they use different rules. Specifically, the rule set contains two different ways of rewriting the category S, both of which result in the same string of words. We will see in Unit 6 that when we have different rules or rule sets that generate the same sentences, we must find ways to decide which rule system represents the best theory.

Sentences that have more than one tree diagram under a given set of rules are said to be **syntactically ambiguous**: the grammar provides more than one way of generating them. Syntactic ambiguity, having more than one tree, is different from **semantic ambiguity**, having more than one meaning. As we will see later on, sentences with more than one tree often do have more than one meaning, but this isn’t always true.

Review

1. Syntax studies what speakers know about the structural arrangement of words and phrases in their language.

The pattern of its forms.



2. Speakers' knowledge of syntax allows them to ...

Construct phrases and sentences out of smaller parts.



3. Phrase structure (PS) rules provide a way of generating sentences. These rules introduce words and tell how those words combine in well-formed strings.

Lexical rules and structural rules.



4. PS rule derivations are conveniently represented with ...

Tree diagrams!



Grammars as Theories

Suppose we have some expressions from a language—a collection of phrases, sentences, and so on. Any set of rules that generates those expressions is called a **grammar** for those expressions.

<p>These sentences were given in Unit 2:</p> <p>Bart ran. Homer sleeps. Maggie crawls. Homer chased Bart. Bart saw Maggie. Maggie petted SLH. Homer handed Lisa Maggie. Marge sent Bart SLH.</p>	<p>This set of rules is a grammar for the sentences:</p> <table style="width: 100%; border: none;"> <tr> <td style="padding-right: 20px;">$S \rightarrow NV$</td> <td>$V \rightarrow \textit{ran}$</td> </tr> <tr> <td style="padding-right: 20px;">$S \rightarrow NVN$</td> <td>$V \rightarrow \textit{sleeps}$</td> </tr> <tr> <td style="padding-right: 20px;">$S \rightarrow NVNN$</td> <td>$V \rightarrow \textit{crawls}$</td> </tr> <tr> <td></td> <td>$V \rightarrow \textit{chased}$</td> </tr> <tr> <td style="padding-right: 20px;">$N \rightarrow \textit{Homer}$</td> <td>$V \rightarrow \textit{saw}$</td> </tr> <tr> <td style="padding-right: 20px;">$N \rightarrow \textit{Marge}$</td> <td>$V \rightarrow \textit{petted}$</td> </tr> <tr> <td style="padding-right: 20px;">$N \rightarrow \textit{Lisa}$</td> <td>$V \rightarrow \textit{sent}$</td> </tr> <tr> <td style="padding-right: 20px;">$N \rightarrow \textit{Bart}$</td> <td>$V \rightarrow \textit{handed}$</td> </tr> <tr> <td style="padding-right: 20px;">$N \rightarrow \textit{Maggie}$</td> <td></td> </tr> <tr> <td style="padding-right: 20px;">$N \rightarrow \textit{Santa's Little Helper}$</td> <td></td> </tr> </table>	$S \rightarrow NV$	$V \rightarrow \textit{ran}$	$S \rightarrow NVN$	$V \rightarrow \textit{sleeps}$	$S \rightarrow NVNN$	$V \rightarrow \textit{crawls}$		$V \rightarrow \textit{chased}$	$N \rightarrow \textit{Homer}$	$V \rightarrow \textit{saw}$	$N \rightarrow \textit{Marge}$	$V \rightarrow \textit{petted}$	$N \rightarrow \textit{Lisa}$	$V \rightarrow \textit{sent}$	$N \rightarrow \textit{Bart}$	$V \rightarrow \textit{handed}$	$N \rightarrow \textit{Maggie}$		$N \rightarrow \textit{Santa's Little Helper}$	
$S \rightarrow NV$	$V \rightarrow \textit{ran}$																				
$S \rightarrow NVN$	$V \rightarrow \textit{sleeps}$																				
$S \rightarrow NVNN$	$V \rightarrow \textit{crawls}$																				
	$V \rightarrow \textit{chased}$																				
$N \rightarrow \textit{Homer}$	$V \rightarrow \textit{saw}$																				
$N \rightarrow \textit{Marge}$	$V \rightarrow \textit{petted}$																				
$N \rightarrow \textit{Lisa}$	$V \rightarrow \textit{sent}$																				
$N \rightarrow \textit{Bart}$	$V \rightarrow \textit{handed}$																				
$N \rightarrow \textit{Maggie}$																					
$N \rightarrow \textit{Santa's Little Helper}$																					

When the set of expressions generated by some rules includes all of the expressions of a language, we'll call the rules a **grammar for the language**.

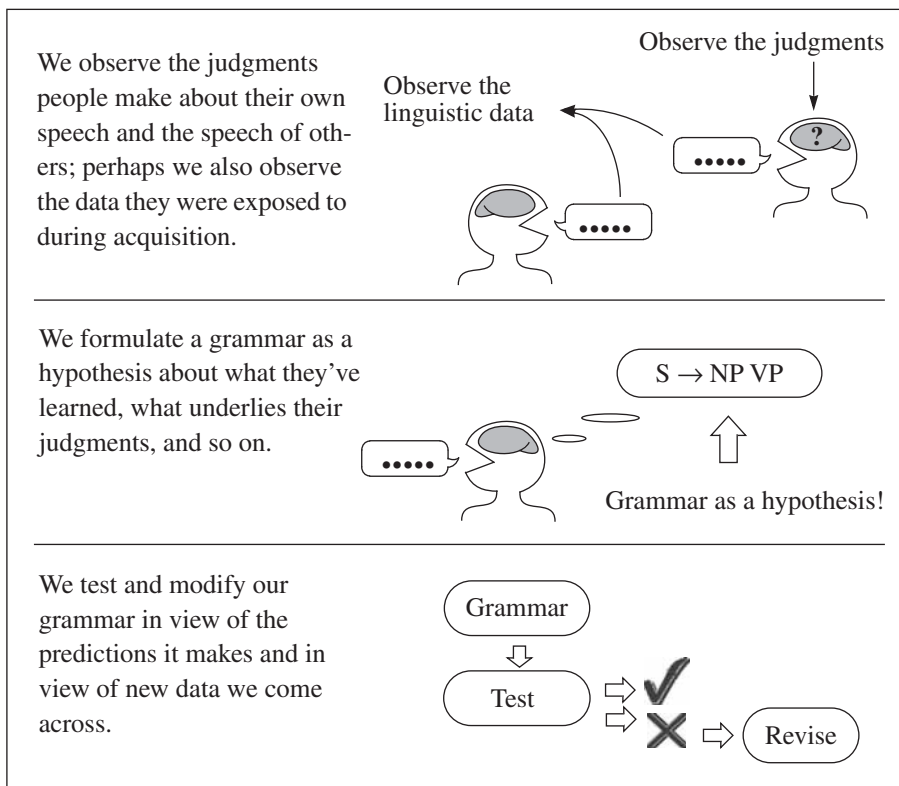
People Know Grammars

The notion of a grammar provides a natural guiding hypothesis about what people know about the syntax of their own language.

Guiding hypothesis

What humans internalize in the course of acquiring their native language is a grammar—a set of rules.

Under this proposal, the grammars that we write down become scientific theories of people's (tacit) syntactic knowledge—candidate solutions to our “black box problem.” As such, they become something to be tested, corrected, refined, and extended, just like any other scientific theory:



We'll look at the third step more closely in the next unit. For the moment, let's look at the first two steps in more detail.

The Data of Syntax

Your knowledge of your native language gives you the ability to judge whether certain strings of words in that language are or are not sentences of the language. Linguists use such well-formedness judgments as a data source in constructing a theory of what speakers know. Native speaker intuitions and judgments are in fact a primary source of data in linguistics.

Judging Well-Formedness Is Not Simple!

Judging well-formedness may seem an easy thing. To determine whether the rule for English sentences is $S \rightarrow N V$ or $S \rightarrow V N$, we just speak sentences with these patterns and listen to whether they sound good or not. What could be simpler? In fact, matters are not so direct.

Well-Formed ≠ Sensible or Natural

Judging whether a sentence of English (or any other language) is well-formed is not the same thing as judging whether it “makes sense” or whether it could ever be used naturally in conversation. Consider examples (1) and (2), due to Chomsky (1957, p. 15):

- (1) Colorless green ideas sleep furiously.
- (2) Revolutionary new ideas happen infrequently.

(1) is clearly nonsensical. We don’t know what it would be like for an idea to be green, never mind for it to be both green and colorless. Likewise, we don’t know what it would mean for ideas to sleep, never mind to sleep furiously. Nonetheless, even though (1) is nonsensical, we recognize it as following an English pattern. (1) has the same grammatical pattern as (2), which is a fully sensible and meaningful sentence of English. In this respect, (1) and (2) contrast sharply with (3), which is not a sentence of English at all:

- (3) Colorless sleep furiously ideas green.

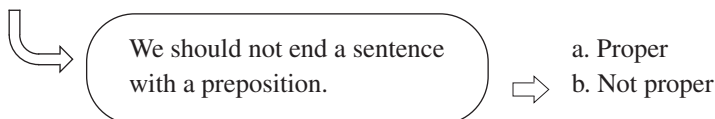
The pattern that we are detecting in (1) and (2) is clearly something independent of what those sentences say or express. It concerns the pure form of these sentences. English speakers know that (1) and (2) share a common formal pattern and that it is a possible pattern for English sentences.

Well-Formed ≠ Proper or Educated

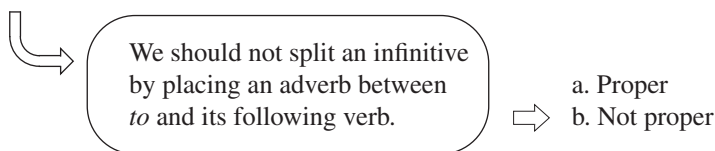
Judging whether a sentence of English (or any other language) is well-formed is also not the same thing as judging whether the sentence sounds “proper” or “correct.” Consider the pairs in (4)–(6):

- (4) a. With whom are you going?
b. Who(m) are you going with?

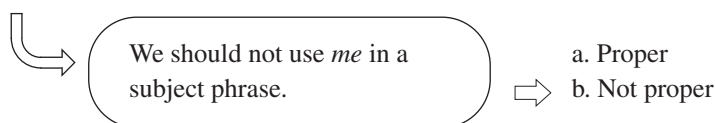
According to rules of “correct” English grammar ...



- (5) a. To go boldly where no one has gone before.
b. To boldly go where no one has gone before.



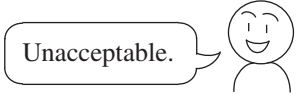

- (6) a. My friend and I just got back from the movies.
b. Me and my friend just got back from the movies.



Even though they may have been taught in school that the patterns in (4b), (5b), and (6b) are “improper” or “incorrect,” many English speakers nonetheless follow these patterns, and understand others who follow them as well. Proper or not, these patterns are part of these speakers’ internal grammar; their linguistic production and comprehension draws on them. As linguists, we are interested in the linguistic patterns that are actually in people’s minds, not in the patterns they are “supposed” to follow but may not. Accordingly, the judgments we are interested in are the ones reflecting the language that people actually speak, not ones reflecting some variant that may be regarded as proper or educated. That is, we are interested in describing the linguistic patterns that speakers actually know, the ones they follow in their own speech. We are not interested in externally prescribed patterns: canons of good English, good French, good Hindi, and so on, that individuals may be aware of and may have been taught in school, but do not really follow. So, again, when we ask others or ourselves whether a given expression is well-formed, we are not asking whether it is grammatically “proper” or “correct.”

Ungrammatical versus Unacceptable

To simplify our discussion, let's adopt a useful terminological distinction. When a speaker rejects a given sentence *for whatever reason*, we'll say that she judges the sentence to be **unacceptable**. When a speaker rejects a sentence because its structural pattern fails to conform to one from her internalized grammar, we'll say that she judges it to be **ungrammatical** or **ill-formed**. Evidently, determining whether a sentence is ungrammatical/ill-formed is much trickier than determining whether it's unacceptable. In the case of unacceptability, we simply ask the speaker whether a sentence is good or not. In the case of ungrammaticality, we must find out whether the unacceptability arises from a particular source. Sentences can be unacceptable for many reasons. Ungrammaticality is a narrower concept.

<p>A sentence rejected by a speaker for whatever reason ...</p> 	<p>A sentence rejected by a speaker because its structural pattern fails to conform to one from her internalized grammar ...</p> 
---	---

Covering the Data


When we go about formulating a grammar, we begin with the judgments that people make about their own speech and about the speech of others. An initial data set might be a list of sentences marked with well-formedness judgments. We looked at a collection of sentences like this in Units 2 and 3. The sentences without stars (asterisks) are ones that English speakers would judge to be grammatical, or **well-formed**. The ones with asterisks are ones that English speakers would judge to be ungrammatical, or **ill-formed**.

I	II	III
Bart ran. Homer sleeps. Maggie crawls. *Ran Maggie. *Crawls Homer.	Homer chased Bart. Bart saw Maggie. Maggie petted SLH. *Chased Bart Homer.	Homer handed Lisa Maggie. Marge sent Bart SLH. *Sent Marge Bart SLH. *Marge Bart SLH sent.

We next formulated a set of rules that generate the unstarred sentences:

S → NV	V → <i>ran</i>
S → NVN	V → <i>sleeps</i>
S → NVNN	V → <i>crawls</i>
	V → <i>chased</i>
N → <i>Homer</i>	V → <i>saw</i>
N → <i>Marge</i>	V → <i>petted</i>
N → <i>Lisa</i>	V → <i>sent</i>
N → <i>Bart</i>	V → <i>handed</i>
N → <i>Maggie</i>	
N → <i>Santa's Little Helper</i>	

Judgments of grammaticality are data. But so are judgments of ungrammaticality. Our theory must cover both!



Notice that although we concentrated on the unstarred sentences, in the sense that those were the ones we aimed at generating, the starred sentences are really just as important. Judgments of grammaticality/well-formedness are data. But so are judgments of ungrammaticality/ill-formedness. Our theory must cover both!

What does that mean, exactly? In what sense can rules cover or account for sentences that aren't grammatical?


In saying that a speaker has internalized a set of syntactic rules, we're claiming that those are the rules the speaker draws on in judging well-formedness and ill-formedness. We're saying that those rules account for the judgments. Accordingly, when we attribute a set of rules to a person, we expect that the person will judge sentences generated by the rules to be well-formed and sentences not generated by the rules to be ill-formed.

Sentences generated by the rules will be judged by the person to be well-formed.

AND

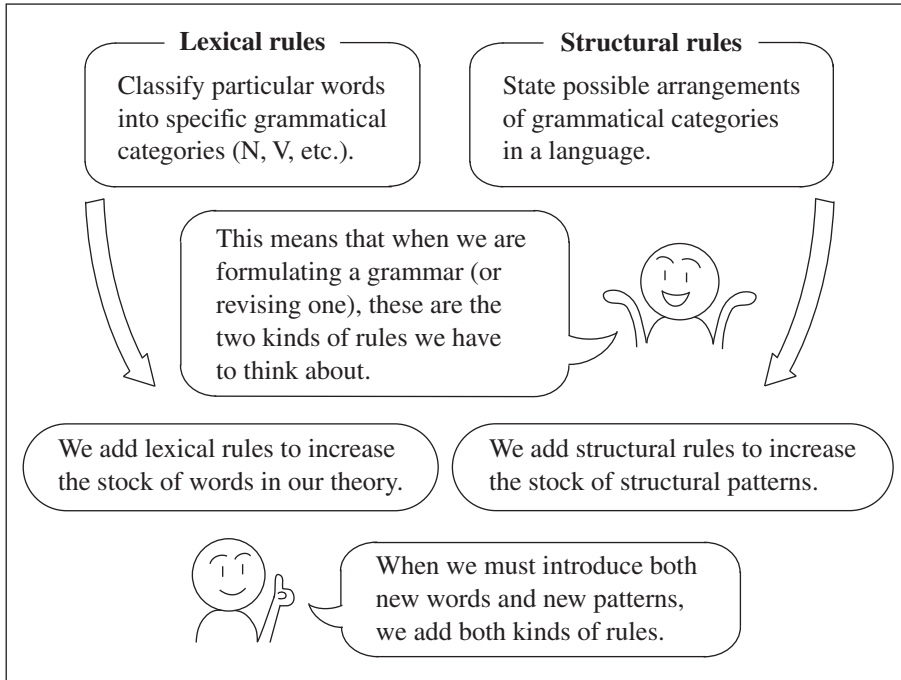
Sentences not generated by the rules will be judged by the person to be ill-formed.

Both kinds of data are relevant and important. Both represent *facts* that we must account for in constructing our theory.



Formulating a Grammar

So far our grammars have been very simple, consisting of just two basic kinds of PS rules:



Building Systematically

When you are trying to formulate a grammar, or any other scientific theory for that matter, there are always a number of ways to proceed. For example, you might simply eyeball a collection of data like this and write down all of the necessary rules in one go:

I	II	III
Bart ran.	Homer chased Bart.	Homer handed Lisa Maggie.
Homer sleeps.	Bart saw Maggie.	Marge sent Bart SLH.
Maggie crawls.	Maggie petted SLH.	*Sent Marge Bart SLH.
*Ran Maggie.	*Chased Bart Homer.	*Marge Bart SLH sent.
*Crawls Homer.		

While this may be reasonable when you are dealing with a few simple pieces of data, it is often useful to follow a more systematic strategy. Here is one “cook-book” recipe for building a grammar:

1. Start with a single piece of data.
2. Build enough of a grammar to account for that one piece of data.
3. Extend your grammar by adding just enough rules to account for the next piece of data.
4. Check to see that your new grammar also accounts for all previous data.
5. Repeat, starting from Step 3.

Notice that when your piece of data is that a certain expression is ill-formed, what you check (then and subsequently) is that your rules *don't* generate this expression!

The idea behind this procedure is simple and obvious. You start with a grammar that covers one fact. You then extend it step by step, always checking to see that, when you add new rules, you haven't lost the ability to generate any well-formed sentences considered previously, and you haven't gained the ability to generate any ill-formed sentences considered previously. This procedure keeps everything under control for you. You build the grammar systematically. For example:



EXAMPLE

Let us apply this procedure to the data in I, II, and III.

First sentence:

Bart ran.

Grammar needed
to generate it:

$S \rightarrow N V$
 $N \rightarrow \textit{Bart}$
 $V \rightarrow \textit{ran}$

Second sentence:

Homer sleeps.

Extend grammar:

$S \rightarrow N V$
 $N \rightarrow \textit{Bart}$
 $N \rightarrow \textit{Homer}$
 $V \rightarrow \textit{ran}$
 $V \rightarrow \textit{sleeps}$



Check that *Bart ran*
is still generated:



Yes, it is!

Third sentence:

Maggie crawls.

Extend grammar:

$S \rightarrow N V$
 $N \rightarrow \textit{Bart}$
 $N \rightarrow \textit{Homer}$
 $N \rightarrow \textit{Maggie}$
 $V \rightarrow \textit{ran}$
 $V \rightarrow \textit{sleeps}$
 $V \rightarrow \textit{crawls}$



Check that *Bart ran*
and *Homer sleeps*
are still generated:



Yes, they are!

Fourth sentence:

*Ran Maggie.

The same grammar:

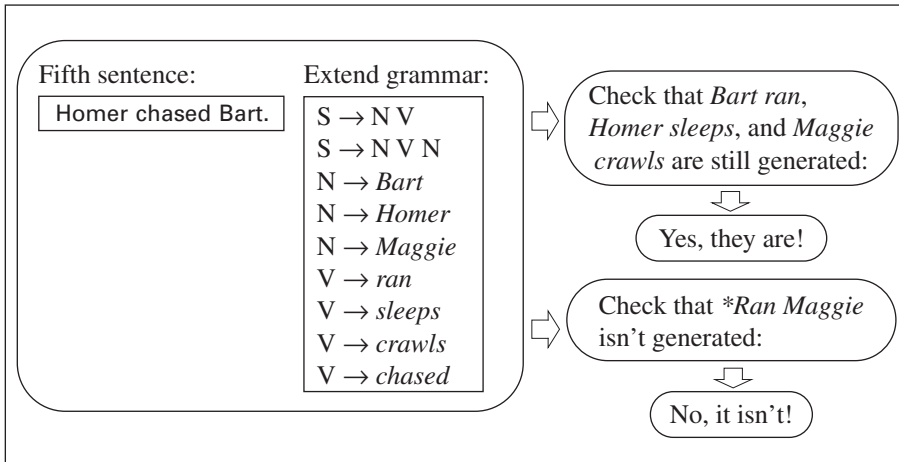
.....







Check that **Ran*
Maggie isn't generated
by the grammar:



No, it isn't!



Review

1. A grammar for some expressions is ...	A set of rules that generate those expressions. 
2. Our guiding hypothesis is that ...	 What people internalize in the course of acquiring their language is a grammar.
3. Grammars are scientific theories of something real, namely ...	The knowledge people have about their language. 
4. Like any other scientific theory, a grammar must be ...	Tested, extended, refined, and revised against the facts. 

Testing a Grammar

Testing a grammar is partly a matter of checking whether its rules generate the expressions you want and don't generate the ones you don't want. For example, suppose we've collected the following data from some individual, Jones:

Data Set 1

Bart ran.	Homer chased Bart.	Homer handed Lisa Maggie.
Homer sleeps.	Bart saw Maggie.	Marge sent Bart SLH.
Maggie crawls.	Maggie petted SLH.	*Sent Marge Bart SLH.
*Ran Maggie.	*Chased Bart Homer.	*Marge Bart SLH sent.
*Crawls Homer.		

Suppose further that we come up with the following grammar:

Grammar A

$S \rightarrow N V$	$N \rightarrow \text{Homer}$	$V \rightarrow \text{ran}$
$S \rightarrow N V N$	$N \rightarrow \text{Marge}$	$V \rightarrow \text{sleeps}$
$S \rightarrow N V N N$	$N \rightarrow \text{Lisa}$	$V \rightarrow \text{crawls}$
	$N \rightarrow \text{Bart}$	$V \rightarrow \text{chased}$
	$N \rightarrow \text{Maggie}$	$V \rightarrow \text{saw}$
	$N \rightarrow \text{Santa's Little Helper}$	$V \rightarrow \text{petted}$
		$V \rightarrow \text{sent}$
		$V \rightarrow \text{handed}$

Part of testing Grammar A will be to check whether its rules generate all the unstarred sentences and none of the starred ones. We do this by attempting derivations for each sentence.

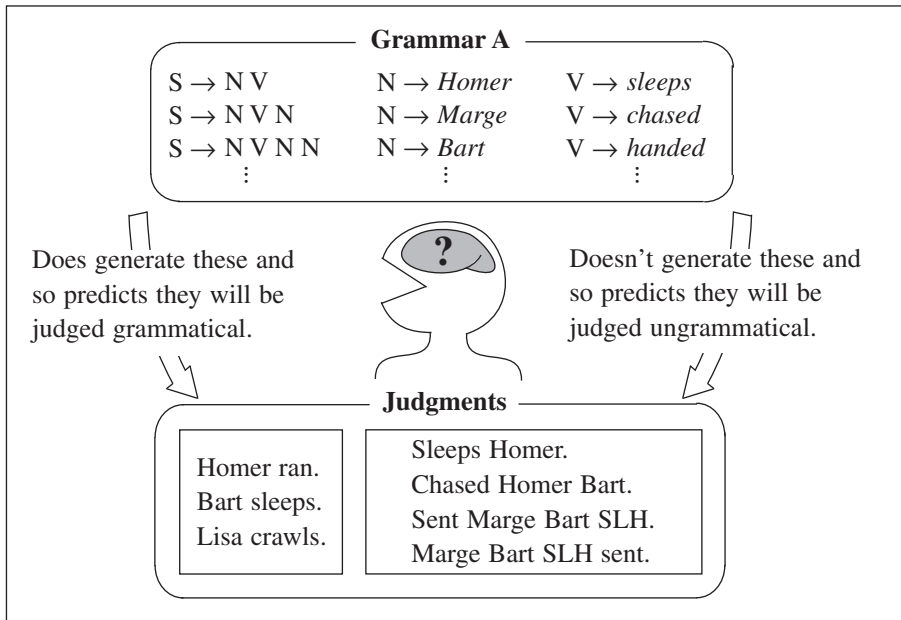
This is not the end of it, however. Notice that Grammar A generates sentences beyond those listed in Data Set 1. It also generates all of the expressions in Data Set 2:

Data Set 2

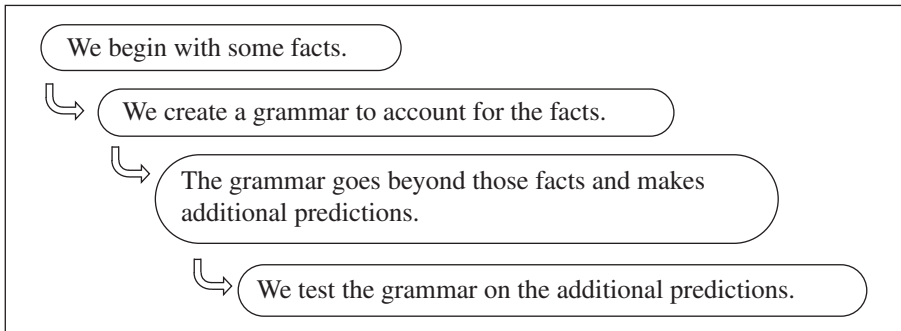
Homer ran.	Bart chased Lisa.
Bart sleeps.	Marge handed Lisa Maggie.
Lisa crawls.	

These sentences are relevant to our testing, too! In claiming that Jones knows Grammar A, we are making the following predictions:

- Sentences generated by Grammar A will be judged to be well-formed by Jones.
- Sentences not generated by Grammar A will be judged to be ill-formed by Jones.



Thus, well-formedness and ill-formedness judgments become **predictions** of the theory. Given that Grammar A generates the additional sentences, we predict that they too will be judged to be well-formed by Jones, even though they go beyond our original data set. They are additional data on which we must test our theory. So the situation is this:



Revising a Grammar

When we test our grammar against additional data, there is of course no guarantee that it will predict correctly. It may well be that the grammar generates sentences that are judged ill-formed by the speaker whose grammar we are trying to model. In this case, we say that our grammar is **incorrect**, or that it **mispredicts** the data.

When a grammar is incorrect, we must **revise** it so as to avoid generating expressions we don't want. Consider Data Set 3 and check whether Grammar A generates these sentences:

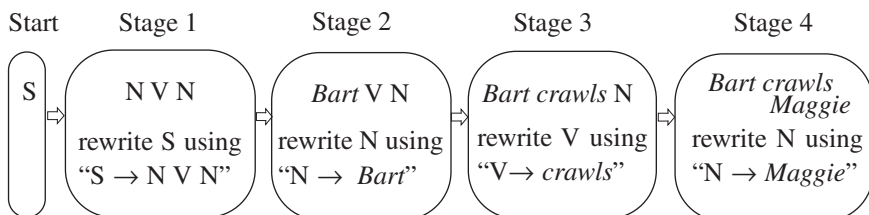
Data Set 3

*Bart crawls Maggie.
*Maggie sleeps Bart.

*Homer ran Bart.
*Maggie handed.

EXAMPLE

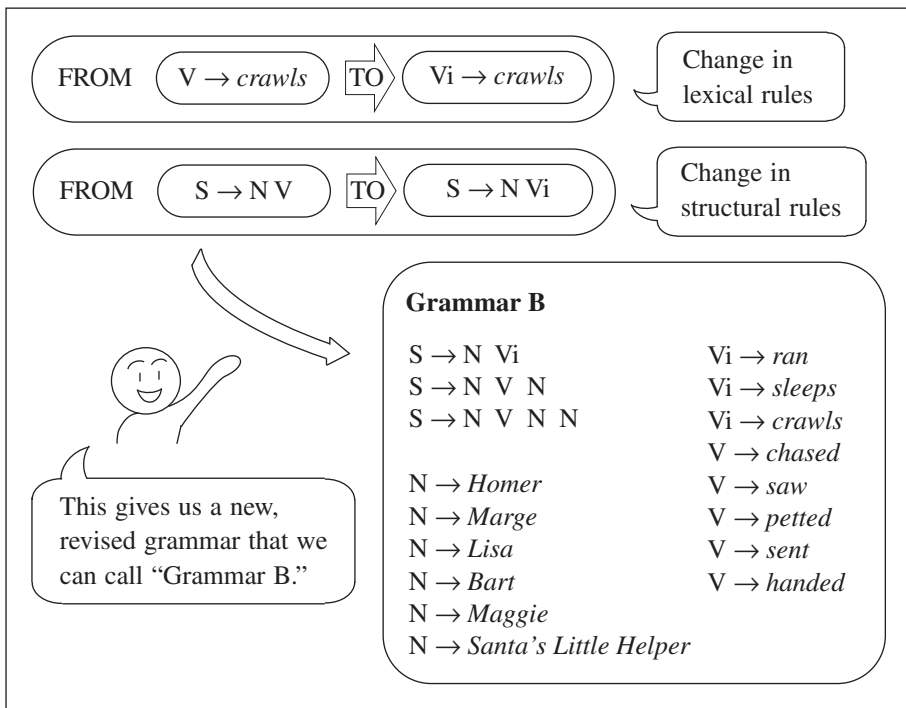
Test Grammar A to see whether it generates the sentence **Bart crawls Maggie*.



It does! But this sentence would be judged by most of us to be ill-formed! Intuitively, *crawls* is not the sort of verb that can be followed by a noun.

Given this result, we must refine Grammar A so that it does not generate this undesirable example (and others like it). How should we do this?

One idea might be to distinguish among types of verbs. That is, we might divide our verbs into two different kinds and assign them to different categories. There would be verbs like *crawls*, which don't take a following noun, and verbs like *saw*, which do take a following noun. Let's call verbs like *crawls* **intransitive verbs** and assign them to their own category V_i . (We'll talk about verbs like *handed* in the next unit.) So we change our lexical rules for intransitive verbs as follows:



If you check, you will see that **Bart crawls Maggie* is no longer generated. We have revised Grammar A so that it no longer mispredicts the data in question.

A Note on Checking Predictions

Remember that when you are checking predictions with speakers, you always have to be careful in evaluating their judgments. Remember that *ill-formed* is different from *senseless*, *unnatural*, *incorrect*, or *improper*. To use terms from Unit 4, *ungrammatical* is different from *unacceptable*. If you present a sentence


to someone and he objects to it, saying, “That sentence sounds bad,” you must be sure about what exactly he is objecting to. Is it the content of the sentence? Its naturalness? Its usefulness? Its status as “proper” English? Its structural form? Only the last constitutes a judgment of ungrammaticality. If the speaker rejects a sentence simply on the grounds that it’s “improper,” you may not want to classify it as ungrammatical.

Extending a Grammar


We’ve seen that a grammar may produce expressions that we don’t want. It may predict expressions to be well-formed that are actually ill-formed. In this case, the grammar is incorrect. A grammar may also fail to generate sentences that we do want. It may fail to predict expressions to be well-formed that are in fact well-formed. In this second case, we say the grammar is **incomplete** or **fails to cover the data adequately**.

When a theory is incomplete, we must **extend** it to generate the expressions that we want. Both Grammars A and B are radically incomplete: they produce only a minuscule part of the full range of English sentences. To develop a grammar that covers anything like the real range, we would have to extend Grammar A or B in at least two ways:

Include more lexical rules


 **EXAMPLE**

Test Grammar B to see whether it generates the sentence *Bart likes Maggie*.


 No, it doesn't.

To generate this sentence, we would need the rule $V \rightarrow \textit{likes}$, which is not in Grammar B.

Include more structural rules

 **EXAMPLE**

Test Grammar B to see whether it generates the sentence *Bart walked to Maggie*.

 No, it doesn't.

To generate this sentence, we would need lexical rules for the words *walked* and *to*, which are not currently in Grammar B. Furthermore, we would need a structural rule to introduce the word *to*, which Grammar B lacks.

We will look at the issues that arise in extending a grammar in much more detail in Unit 6.

EXERCISES

1. Consider the sentence *Homer saw her duck*. It has two meanings, which correspond to two different sentence patterns. What are the two patterns?

2. Here is a set of phrase structure rules for English:

$S \rightarrow N V$	$V \rightarrow \textit{ran}$
$S \rightarrow N V N$	$V \rightarrow \textit{saw}$
$S \rightarrow N V N N$	$V \rightarrow \textit{sleeps}$
	$V \rightarrow \textit{fed}$
$N \rightarrow \textit{Homer}$	$V \rightarrow \textit{crawls}$
$N \rightarrow \textit{Marge}$	$V \rightarrow \textit{gave}$
$N \rightarrow \textit{Lisa}$	$V \rightarrow \textit{chased}$
$N \rightarrow \textit{Bart}$	$V \rightarrow \textit{sent}$
$N \rightarrow \textit{Maggie}$	
$N \rightarrow \textit{SLH}$	

These rules generate the sentences in (1):

(1) Bart ran. Homer chased Bart. Marge gave Homer Maggie.
Homer sleeps. Lisa saw Maggie. Homer sent Bart SLH.
Maggie crawls. Maggie fed SLH.

A. What tree diagram do the rules give for the sentence *Maggie fed SLH*?

B. Give four other sentences of English that these rules generate (i.e., find examples different from the ones in (1)).

3. The sentences below show new patterns, different from the ones in (1) of Question 2:

(1) Homer talked to Marge.
Homer talked about Bart.
Maggie crawled to Lisa.
SLH ran from Homer.
Homer talked to Marge about Bart.
Maggie crawled from Lisa to Marge.

A. What new rules must be added to the rules in Question 2 in order to produce these sentences?

- B. What tree diagram do your new rules give for the sentence *Homer talked to Marge about Bart*?
4. The sentences in (1) show yet another sentence pattern, different from the ones in Questions 2 and 3.
- (1) Homer talked to Bart yesterday.
Marge gave Homer Maggie quickly.
Homer chased Bart recently.
- A. What new rules must be added in order to produce these sentences?
- B. What tree diagrams do your new rules give for the sentences *Homer talked to Bart yesterday* and *Homer chased Bart recently*?
5. *Bart chased Lisa* is a sentence (S) with the pattern N V N. Now consider the sentence *Marge thinks Bart chased Lisa*. One way to state the pattern of this sentence is N V N V N. But there's a better way. What is it?
6. Below is a set of phrase structure rules for English. (Ignore what *CN* and *Art* stand for.)

S → NP V NP	N → <i>Bart</i>
S → NP V NP NP	N → <i>Marge</i>
NP → Art CN	N → <i>Homer</i>
NP → NP <i>and</i> NP	N → <i>Lisa</i>
NP → N	V → <i>bought</i>
	V → <i>saw</i>
Art → <i>a</i>	V → <i>sent</i>
CN → <i>beer</i>	
CN → <i>gift</i>	

- A. These rules generate a tree for the sentence *Homer bought Marge a gift*. Give the tree.
- B. These rules generate a tree for the sentence *Homer sent Marge Bart and Lisa*. Give the tree.
5. Below is a grammar for a small part of English. (Again, ignore what the new category symbols may stand for.)

S → NP V NP	Art → <i>the</i>
S → NP V NP PP	CN → <i>man</i>
S → NP V NP AP	CN → <i>vase</i>
NP → Art CN	CN → <i>judge</i>
NP → NP AP	N → <i>Homer</i>
NP → N	N → <i>Marge</i>
AP → A	V → <i>considers</i>

$V \rightarrow \textit{found}$
 $A \rightarrow \textit{intelligent}$
 $A \rightarrow \textit{broken}$
 $A \rightarrow \textit{guilty}$
 $PP \rightarrow \textit{there}$

- A. This grammar generates a tree for the sentence *The man found Homer there*. Give the tree.
- B. This grammar assigns two different trees to the sentence *Marge found the vase broken* (that is, the sentence is syntactically ambiguous under these rules). Give the two trees.
- C. To generate the sentences in (1)–(3), you must add a rule or some rules to the grammar. State what rule(s) you must add. (Note: Think of this as a cumulative process, so for each sentence, list only a rule or rules that you haven't added at an earlier point.)

- (1) The man arrived tired.
- (2) a. A tall man arrived.
b. Marge saw a tall man.
- (3) Bart left the party angry at Lisa.

6. Here is a grammar for a small part of English:

$S \rightarrow NP V NP$	$N \rightarrow \textit{Homer}$
$S \rightarrow NP V PP$	$N \rightarrow \textit{Marge}$
$S \rightarrow NP V NP P$	$V \rightarrow \textit{decided}$
$NP \rightarrow \textit{Art CN}$	$V \rightarrow \textit{considered}$
$NP \rightarrow N$	$V \rightarrow \textit{looked}$
$V \rightarrow V P$	$P \rightarrow \textit{up}$
$PP \rightarrow P NP$	$P \rightarrow \textit{on}$

$\textit{Art} \rightarrow \textit{the}$
 $\textit{CN} \rightarrow \textit{answer}$
 $\textit{CN} \rightarrow \textit{boat}$
 $\textit{CN} \rightarrow \textit{present}$

- A. This grammar generates a tree for the sentence *Homer looked Marge up*. Give the tree.
- B. This grammar assigns two different trees to the sentence *Marge decided on the boat* (that is, the sentence is syntactically ambiguous under these rules). Give the two trees.

C. To generate the sentences in (1)–(2), you must add a rule or some rules to the grammar. State what rule(s) you must add. (Note: Think of this as a cumulative process, so for each sentence, list only a rule or rules that you haven't added at an earlier point.)

(1) Homer looked Bart over.

(2) Marge looked the new answer up.

7. Here is a grammar for a small part of English:

S → NP Vi	Art → <i>a</i>
S → NP Vd NP PP	Art → <i>the</i>
S → NP Vt NP	Vi → <i>ran</i>
NP → N	Vi → <i>slept</i>
NP → Art CN	Vi → <i>crawled</i>
PP → P NP	Vt → <i>chased</i>
	Vt → <i>saw</i>
N → <i>Homer</i>	Vt → <i>knew</i>
N → <i>Maggie</i>	Vd → <i>gave</i>
N → <i>Marge</i>	Vd → <i>sent</i>
N → <i>Lisa</i>	P → <i>to</i>
N → <i>Bart</i>	
CN → <i>man</i>	
CN → <i>woman</i>	
CN → <i>girl</i>	
CN → <i>boy</i>	

Extending the grammar

Now, here is a list of sentences:

- | | |
|--------------------------------------|----------------------------------|
| (1) Maggie left the room after Lisa. | Marge told Lisa about Bart. |
| Marge put a hat on Bart. | Marge wrote a letter to Bart. |
| Homer put Maggie near Lisa. | Marge wrote a letter about Bart. |
| Bart put Maggie in the crib. | Marge wrote a letter about Bart. |

- A. State what new rules must be added to the grammar in order to generate the sentences in (1).
- B. Give the tree diagrams that the grammar plus your new rules assign to these eight sentences. (You will need eight trees.)

Testing and revising the grammar

- C. State whether your new, amended grammar generates the following sentences:
- (2) *Marge put a hat to Bart. *Marge gave a hat near Bart.
 *Marge gave a hat on Bart. *Marge told Lisa to Bart.
 Marge wrote a letter near Bart. Marge wrote a letter after Bart.
 Marge wrote a letter on Bart.
- D. If your rules generate any of the ill-formed sentences in (2), revise them so that they do not.

Evaluating additional data

Consider the following additional examples:

- (3) Marge wrote. Marge wrote a letter. Marge wrote to Bart.
 Marge gave. Marge gave money. Marge gave to charity.
 *Marge put. *Marge put the hat. *Marge put on charity.
- E. What further questions do these facts raise for your rules?
- F. What analysis should be given for them?