1

THE INFINITE LIBRARY

Language comes so naturally to us that it is easy to forget what a strange and miraculous gift it is. All over the world members of our species fashion their breath into hisses and hums and squeaks and pops and listen to others do the same. We do this, of course, not only because we like the sounds but because details of the sounds contain information about the intentions of the person making them. We humans are fitted with a means of sharing our ideas, in all their unfathomable vastness. When we listen to speech, we can be led to think thoughts that have never been thought before and that never would have occurred to us on our own. Behold, the bush burned with fire, and the bush was not consumed. Man is born free, and everywhere he is in chains. Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence. Energy equals mass times the speed of light squared. I have found it impossible to carry the heavy burden of responsibility and to discharge my duties as King without the help and support of the woman I love.

Language has fascinated people for thousands of years, and linguists have studied every detail, from the number of languages spoken in New Guinea to why we say *razzle-dazzle* instead of *dazzle-razzle*. Yet to me the first and deepest challenge in understanding language is accounting for its boundless expressive power. What is the trick behind our ability to fill one another's heads with so many different ideas?

The premise of this book is that there are two tricks, words and rules. They work by different principles, are learned and used in different ways, and may even reside in different parts of the brain. Their border disputes shape and reshape languages over centuries, and make language not only a tool for communication but also a medium for wordplay and poetry and an heirloom of endless fascination.

~

The first trick, the word, is based on a memorized arbitrary pairing between a sound and a meaning. "What's in a name?" asks Juliet. "That which we call a rose by any other name would smell as sweet." What's in a name is that everyone in a language community tacitly agrees to use a particular sound to convey a particular idea. Although the word *rose* does not smell sweet or have thorns, we can use it to convey the idea of a rose because all of us have learned, at our mother's knee or in the playground, the same link between a noise and a thought. Now any of us can convey the thought by making the noise.

The theory that words work by a conventional pairing of sound and meaning is not banal or uncontroversial. In the earliest surviving debate on linguistics, Plato has Hermogenes say, "Nothing has its name by nature, but only by usage and custom." Cratylus disagrees: "There is a correctness of name existing by nature for everything: a name is not simply that which a number of people jointly agree to call a thing." Cratylus is a creationist, and suggests that "a power greater than man assigned the first names to things." Today, those who see a correctness of names might attribute it instead to onomatopoeia (words such as *crash* and *oink* that sound like what they mean) or to sound symbolism (words such as *sneer*, *cantankerous*, and *mellifluous* that naturally call to mind the things they mean).

Today this debate has been resolved in favor of Hermogenes' conventional pairing. Early in this century Ferdinand de Saussure, a founder of modern linguistics, called such pairing the *arbitrary sign* and made it a cornerstone of the study of language.[1] Onomatopoeia and sound symbolism certainly exist, but they are asterisks to the far more important principle of the arbitrary sign—or else we would understand the words in every foreign language instinctively, and never need a dictionary for our own! Even the most obviously onomatopoeic words—those for animal sounds—are notoriously unpredictable, with pigs oinking *boo-boo* in Japan and dogs barking *gong-gong* in Indonesia. Sound symbolism, for its part, was no friend of the American woman in the throes of labor who overheard what struck her as the most beautiful word in the English language and named her newborn daughter *Meconium*, the medical term for fetal excrement.[2]

Though simple, the principle of the arbitrary sign is a powerful tool for getting thoughts from head to head. Children begin to learn words before their first birthday, and by their second they hoover them up at a rate of one every two hours. By the time they enter school children command 13,000 words, and then the pace picks up, because new words rain down on them from both speech and print. A typical high-school graduate knows about 60,000 words; a literate adult, perhaps twice that number.[3] People recognize words swiftly. The meaning of a spoken word is accessed by a listener's brain in about a fifth of a second, before the speaker has finished pronouncing it.[4] The meaning of a printed word is registered even more quickly, in about an eighth of a second.[5] People produce words almost as rapidly: It takes the brain about a quarter of a second to find a word to name an object, and about another quarter of a second to program the mouth and tongue to pronounce it.[6]

The arbitrary sign works because a speaker and a listener can call on identical entries in their mental dictionaries. The speaker has a thought, makes a sound, and counts on the listener to hear the sound and recover that thought. To depict an entry in the mental dictionary we need a way of showing the entry itself, as well as its sound and meaning. The entry for a word is simply its address in one's memory, like the location of the boldfaced entry for a word in a real dictionary. It's convenient to use an English letter sequence such as *r-o-s-e* to stand for the entry, as long as we remember this is just a mnemonic tag that allows us to remember which word the entry corresponds to; any symbol, such as 42759, would do just as well. To depict the word's sound, we can use a phonetic notation, such as [rōz].* The meaning of a word is a link to an entry in the person's mental *encyclopedia*, which captures the person's concept of a rose. For convenience we can symbolize it with a picture, such as ❀. So a mental dictionary entry looks something like this:

rose
  sound: rōz
  meaning: ❀

---

*This book uses a simplified phonetic notation similar to that found in dictionaries, in which the long vowels ā in *bait*, ē in *beet*, ī in *bite*, ō in *boat*, and ū in *boot* are distinguished from the short vowels ă in *bat*, ĕ in *bet*, ĭ in *bit*, ŏ in *pot*, and ŭ in *but*. An unadorned *a* stands for the first vowel in *father* or *papa*. The symbol ɨ is used for the neutral vowel in the suffix of *melted* and *Rose's* (e.g., ***mĕltɨd***, ***rōzɨz***), a version of the vowel sometimes called *schwa*.

"Long vowel," "short vowel," and other technical terms in linguistics, psycholinguistics, and neuroscience are defined in the Glossary.

A final component is the word's part of speech, or grammatical category, which for *rose* is noun (N):

rose
  sound: *rōz*
  meaning: ❀
  part of speech: N

And that brings us to the second trick behind the vast expressive power of language.

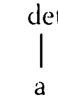
~

People do not just blurt out isolated words but rather *combine* them into phrases and sentences, in which the meaning of the combination can be inferred from the meanings of the words and the way they are arranged. We talk not merely of roses, but of the red rose, proud rose, sad rose of all my days. We can express our feelings about bread and roses, guns and roses, the War of the Roses, or days of wine and roses. We can say that lovely is the rose, roses are red, or a rose is a rose is a rose. When we combine words, their arrangement is crucial: *Violets are red, roses are blue*, though containing all the ingredients of the familiar verse, means something very different. We all know the difference between *young women looking for husbands* and *husbands looking for young women*, and that *looking women husbands young for* doesn't mean anything at all.

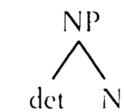
Inside everyone's head there must be a code or protocol or set of rules that specifies how words may be arranged into meaningful combinations. Modern linguists call it a *grammar*, sometimes a *generative grammar* to distinguish it from the grammars used to teach foreign languages or to teach the dos and don'ts of formal prose.

A grammar assembles words into phrases according to the words' part-of-speech categories, such as noun and verb. To highlight a word's category and reduce visual clutter often it is convenient to omit the sound and meaning and put the category label on top:

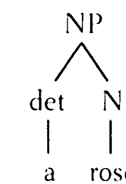
Similarly, the word *a*, an article or *determiner*, would look like this:



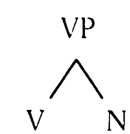
They can then be joined into the phrase *a rose* by a rule that joins a determiner to a noun to yield a noun phrase (NP). The rule can be shown as a set of connected branches; this one says "a noun phrase may be composed of a determiner followed by a noun":



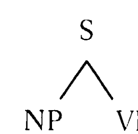
The symbols at the bottom of the branches are like slots into which words may be plugged, as long as the words have the same labels growing out of their tops. Here is the result, the phrase *a rose*:



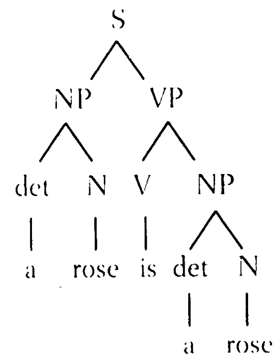
With just two more rules we can build a complete toy grammar. One rule defines a predicate or verb phrase (VP); the rule says that a verb phrase may consist of a verb followed by its direct object, a noun phrase:



The other rule defines the sentence itself (S). This rule says that a sentence may be composed from a noun phrase (the subject) followed by a verb phrase (the predicate):



When words are plugged into phrases according to these rules, and the phrases are plugged into bigger phrases, we get a complete sentence, such as *A rose is a rose*:



Other parts of the rules, not shown here, specify the meaning of the new combination. For example, the complete NP rule says that the meaning of *the yellow rose of Texas* is based on the meaning of *rose*, which is called the *head* of the phrase, and that the other words modify the head in various ways: *yellow* specifies a distinctive trait, *Texas* its location.

These rules, though crude, illustrate the fantastic expressive power made available by grammar. First, the rules are *productive*. By specifying a string of *kinds* of words rather than a string of *actual* words, the rules allow us to assemble new sentences on the fly and not regurgitate preassembled clichés—and that allows us to convey unprecedented combinations of ideas. Though we often speak of roses being red, we could talk about violets being red if the desire came over us (perhaps to announce a new hybrid), because the rule allows us to insert *violets* into the N slot just as easily as *roses*.

Second, the symbols contained by the rules are *symbolic* and hence *abstract*. The rule doesn't say, "A sentence may begin with a bunch of words referring to a kind of flower"; rather, it says, "A sentence may begin with an NP," where NP is a symbol or variable that can be replaced by any noun, just as *x* or *y* in a mathematical formula can be replaced by any number. We can use the rules to talk about flowers and their colors and smells, but we can just as easily use them to talk about karma or quarks or floob-boober-bab-boober-bubs (who, according to Dr. Seuss, bounce in the water like blubbery tubs).

Third, the rules are *combinatorial*. They don't just have a single slot, like a fill-in-the-blank exam question; *every* position in the sentence offers a choice of words from a lengthy menu. Say everyday English has four determiners (*a*, *any*, *one*, and *the*) and ten thousand nouns. Then the rule for a noun phrase allows four choices for the determiner, followed by ten thousand choices for the head noun, yielding 4 × 10,000 = 40,000 ways to utter a noun phrase. The rule for a sentence allows these forty thousand subjects to be followed by any of four thousand verbs, providing 40,000 × 4,000 = 160,000,000 ways to utter the first three words of a sentence. Then there are four choices for the determiner of the object (640 million four-word beginnings) followed by ten thousand choices for the head noun of the object, or 640,000,000 × 10,000 = 6,400,000,000,000 (6.4 trillion) five-word sentences. Suppose it takes five seconds to produce one of these sentences. To crank them all out, from *The abandonment abased the abbey* and *The abandonment abased the abbot*, through *The abandonment abased the zoologist*, all the way to *The zoologist zoned the zoo*, would take a million years.

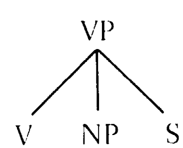
Many such combinations are ungrammatical of course, owing to various complications I haven't mentioned—for example, you can't say *The Aaron, a abandonment*, or *The abbot abase the abbey*. And most of the combinations are nonsensical: Abandonments can't abbreviate, and abbeys can't abet. Yet even with these restrictions the expressive range of a grammar is astonishing. The psychologist George Miller once conservatively estimated that if speakers keep a sentence perfectly grammatical and sensible as they choose their words, their menu at each point offers an average of about ten choices (at some points there are many more than ten choices; at others, only one or two).[7] That works out to one hundred thousand five-word sentences, one million six-word sentences, ten million seven-word sentences, and so on. A sentence of twenty words is not at all uncommon (the preceding sentence has twenty words before *and so on*), and there are about one hundred million trillion of them in English. For comparison, that is about a hundred times the number of seconds since the birth of the universe.

Grammar is an example of a combinatorial system, in which a small inventory of elements can be assembled by rules into an immense set of distinct objects. Combinatorial systems obey what Miller calls the Exponential Principle: The number of possible combinations grows exponentially (geometrically) with the size of the combination.[8] Combinatorial systems can generate inconceivably vast numbers of products. Every kind of molecule in the universe is assembled from a hundred-odd chemical elements; every protein building block and catalyst in the living world is assembled from just twenty amino acids. Even when the number of products is smaller, a combinatorial system can capture them all and provide enormous savings in storage space. Eight bits define $2^8 = 256$ distinct bytes, which is more than enough for all the numerals, punctuation marks, and upper- and lowercase letters in our writing system. This allows computers to be built out of identical specks of silicon that can be in just two states, instead of the dozens of pieces of type that once filled typesetters' cases. Billions of years ago life on Earth settled on a code in which a string of three bases in a DNA molecule became the instruction for selecting one amino acid when assembling a protein. There are four kinds of bases, so a three-base string allows for $4 \times 4 \times 4 = 64$ possibilities. That is enough to give each of the twenty amino acids its own string, with plenty left over for the start and stop instructions that begin and end the protein. Two bases would have been too few ($4 \times 4 = 16$), four more than needed ($4 \times 4 \times 4 \times 4 = 256$).

Perhaps the most vivid description of the staggering power of a combinatorial system is in Jorge Luis Borges's story "The Library of Babel."[9] The library is a vast network of galleries with books composed of all the combinations of twenty-two letters, the comma, the period, and the space. Somewhere in the library is a book that contains the true history of the future (including the story of your death), a book of prophecy that vindicates the acts of every man in the universe, and a book containing the clarification of the mysteries of humanity. People roamed the galleries in a futile search for those texts from among the untold number of books with false versions of each revelation, the millions of facsimiles of a given book differing by a character, and, of course, the miles and miles of gibberish. The narrator notes that even when the human species goes extinct, the library, that space of combinatorial possibilities, will endure: "illuminated, solitary, infinite, perfectly motionless, equipped with precious volumes, useless, incorruptible, secret."

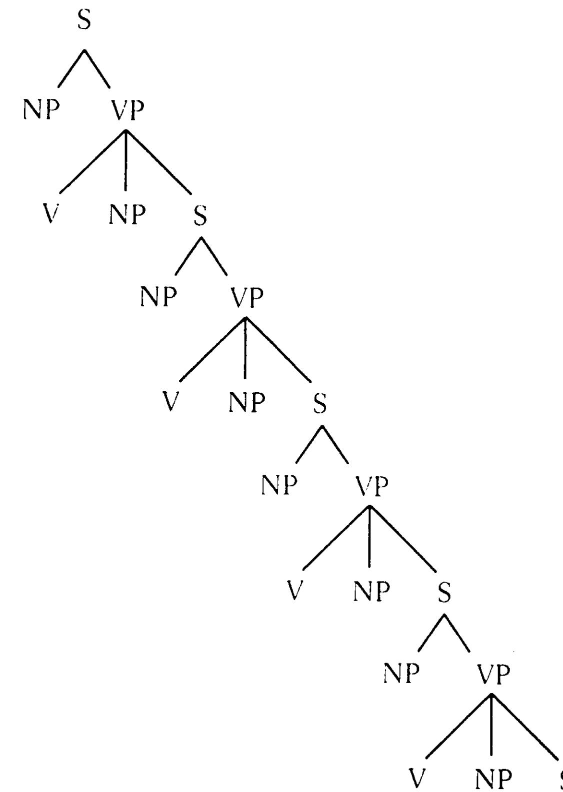
Technically, Borges needn't have described the library as "infinite." At eighty characters a line, forty lines a page, and 410 pages a book, the number of books is around $10^{1,800,000}$, or 1 followed by 1.8 million zeroes. That is, to be sure, a very large number—there are only 10^{70} particles in the visible universe—but it is a finite number.

It is easy to make a toy grammar that is even more powerful than the scheme that generates The Library of Babel. Suppose our rule for the verb phrase is enriched to allow a sentence (S) to appear inside it, as in *I told Mary he was a fool*, in which *he was a fool* comes after the object NP *Mary*:



Now our grammar is *recursive*: The rules create an entity that can contain an example of itself. In this case, a Sentence contains a Verb Phrase which in

turn can contain a Sentence. An entity that contains an example of itself can just as easily contain an example of itself that contains an example of itself that contains an example of itself, and so on:



In this case a sentence can contain a verb phrase, which can contain a sentence, which can contain a verb phrase, which can contain a sentence, ad infinitum. For example, I think I'll tell you that I just read a news story that recounts that Stephen Brill reported that the press uncritically believed Kenneth Starr's announcement that Linda Tripp testified to him that Monica Lewinsky told Tripp that Bill Clinton told Vernon Jordan to advise Lewinsky not to testify to Starr that she had had a sexual relationship with Clinton. That statement is a Russian doll with thirteen sentences inside sentences inside sentences. A recursive grammar can generate sentences of any length, and thus can generate an infinite number of sentences. So a human being possessing a recursive grammar can express or understand an infinite number of distinct thoughts, limited in practice only by stamina and mortality.

The idea that the creativity inherent in language can be explained by a grammar of combinatorial rules is usually associated with the linguist Noam Chomsky. Chomsky traced the idea to Wilhelm von Humboldt, a nineteenth-century pioneer of linguistics, who explained language as "the infinite use of finite media." According to Chomsky, the idea is even older than that; Humboldt was the last in a tradition of "Cartesian linguists" dating back to the Enlightenment.[10]

Enlightenment philosophers were captivated by the dizzying range of thoughts made expressible by a combinatorial grammar. In his book *The Search for the Perfect Language* the semiotician Umberto Eco recounts the many Promethean schemes these philosophers came up with to perfect and harness their power.[11] Descartes noticed that the decimal system allows a person to learn in a day the names of all the quantities to infinity, and he suggested that a universal artificial language built on similar principles could organize all human thoughts. Leibniz, too, dreamed of a universal logical grammar that would generate only valid sequences of ideas, banishing irrationality and error forever.

Three hundred years later we still are fallible, and still take years to learn a Babel of local languages with their tens of thousands of arbitrary signs. Why has no modern language used the horsepower of combinatorial grammar to the fullest and abandoned the unprincipled, parochial, onerous-to-memorize laundry list called vocabulary? The answer becomes clear when we look at the most famous of the combinatorial schemes of the Enlightenment, the philosophical language of Bishop John Wilkins. The arbitrary name was an affront to Wilkins's sense of good design, and he strove for a way to eliminate it. He wrote, "We should, by learning . . . the *Names* of things, be instructed likewise in their *Natures*."

Wilkins's system, laid out in a lengthy 1668 opus, offered the user a *nonarbitrary* name for every thing by dividing the universe into categories and subcategories and sub-subcategories, and assigning a vowel or consonant to every branch in the tree. The first syllable identified one of the forty categories into which Wilkins had sorted all thinkable thoughts. For example, Z stood for "sensitive species" (animals) and could be followed by *i* for "beasts" (quadrupeds). The next consonant picked out a subdivision; *t*, for example, stood for rapacious terrestrial European canines. A final vowel pinpointed the species, yielding *Zita* as the name for dogs. By similar computations one

could deduce another two thousand names for things. *Zana* is a scaly river fish with reddish flesh, in other words, salmon. *Siba* is a type of public military relation, namely, defense. *Deba* is a portion of the first of the terrestrial elements (fire), to wit, flame. *Coba* is a consanguinous economic relation of direct ascendant, a.k.a. father.

Wilkins's philosophical language has been analyzed insightfully by Borges and Eco, and we can see why no one today speaks Wilkish.[12] For one thing, it forces users to perform a chain of computations in their heads every time they want to refer to a dog. Every vowel and consonant is laden with meaning and acts as a premise in a lengthening deduction. Speakers of the language would have to play a game of Twenty Questions, inferring an entity from a description, for every word in a sentence. They could of course simply memorize the answers, such as that a portion of the first of the terrestrial elements is a flame, but that is not much easier than memorizing that the word for flame is *flame*.

A second problem is that there are more things in heaven and earth than were dreamt of in Wilkins's philosophy, which identified only two thousand concepts. Wilkins understood the exponential principle and tried to cope with the problem by lengthening the words. He provided suffixes and connectors that allowed *calf*, for example, to be expressed as *cow* + *young*, and *astronomer* to be expressed as *artist* + *star*. But eventually he gave up and resorted to using synonyms for concepts his language could not generate, such as *box* for *coffin*. Wilkins's dilemma was that he could either expand his system to embrace all concepts, which would require even longer and more unwieldy strings, or he could force his users to remember the nearest synonym, reintroducing the despised memorization process.

A third problem is that in a logical language words are assembled purely on information-theoretic principles, with no regard to the problems that incarnate creatures might have in pronouncing and understanding the strings. A perfect combinatorial language is always in danger of generating mouthfuls like *mxyzptlk* or *bftsplk*, so Wilkins and other language-designers of the Enlightenment all had to make concessions to pronounceability and euphony. Sometimes they defiled their systems with irregularities, for example, reversing a vowel and consonant to make a word more pronounceable. At other times they hobbled the system with restrictions, such as that consonants and vowels must alternate. Every even-numbered position in a word had to be filled by one of the nine vowels of English, and that restricted many categories, such as species in a genus, to nine apiece, regardless of how many species exist in the world.

Another problem is that Wilkins's words are packed tight with information and lack the safety factor provided by redundancy. The slightest slip of the tongue or pen guarantees misunderstanding. Eco catches Wilkins himself misusing *Gade* (barley) for *Gape* (tulip).

Finally, all that power is not being put to any sensible use. The beauty of a combinatorial system is that it generates combinations that have never before been considered but that one *might* want to talk about some day. For example, the combinatorial system known as the periodic table of the elements inspired chemists to look for hitherto unknown chemical elements that should have occupied the empty slots in the table. Combinatorial grammar allows us to talk about a combinatorial world, a world in which violets could be red or a man could bite a dog. Yet familiar objects and actions around us often form a *non*combinatorial list of distinctive kinds. When we merely have to single out one of them, a combinatorial system is overkill. We never will have to refer to fish with an enmity to sheep or to military actions with scales and reddish flesh, and that's what a combinatorial system for words like Wilkins's allows us to do. To refer to everyday things it's easier to say *dog* or *fish* than to work through a complicated taxonomy that is just a fancy way of singling out dogs or fish anyway.

~

The languages of Wilkins and other Enlightenment thinkers show that combinatorial grammar has disadvantages as well as advantages, and that illuminates our understanding of the design of human language. No language works like Wilkins's contraption, with every word compiled out of meaningful vowels and consonants according to a master formula. All languages force their speakers to memorize thousands of arbitrary words, and now we can see why.[13] Many bodily organ systems are made from several kinds of tissue optimized for jobs with contradictory specifications. Our eyes have rods for night vision and cones for day vision; our muscles have slow-twitch fibers for sustained action and fast-twitch fibers for bursts of speed. The human language system also appears to be built out of two kinds of mental tissue. It has a lexicon of words, which refer to common things such as people, places, objects, and actions, and which are handled by a mechanism for storing and retrieving items in memory. And it has a grammar of rules, which refer to novel relationships

among things, and which is handled by a mechanism for combining and analyzing sequences of symbols.

To a parsimonious scientific mind, however, two mental mechanisms can be one too many. The poet William Empson wrote of the Latin philosopher,

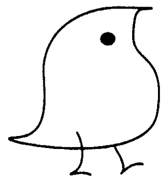
> Lucretius could not credit centaurs;
> Such bicycle he deemed asynchronous.[14]

Today's skeptics also might wonder about a two-part design for language. Perhaps words and rules are two modes of operation of a single faculty. Simple, familiar thoughts need short noises, which we call words, and complicated, unfamiliar thoughts need long noises, which we call phrases and sentences. A single machine might make either short or long noises, depending on the kinds of thoughts it is asked to express. Or perhaps there is a gradual continuum between memory and combination rather than two distinct mechanisms, with words at the memory end of the continuum and sentences at the combination end.

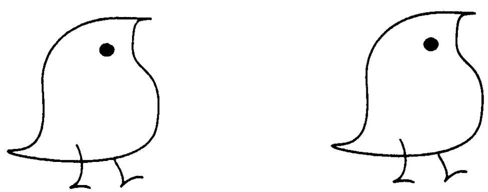
To show that words and rules are handled by different machines we need to hold the input and output of the putative machines constant. We need side-by-side specimens in which the same kind of thought is packed into the same kind of verbiage, but one specimen shows the handiwork of a word regurgitator and the other shows the handiwork of a rule amalgamator. I believe that languages do provide us with such specimens. They are called regular and irregular words.

English verbs come in two flavors. Regular verbs have past tense forms that look like the verb with *-ed* on the end: Today I *jog*, yesterday I *jogged*. They are monotonously predictable: *jog–jogged, walk–walked, play–played, kiss–kissed*, and so on. (Regular nouns, whose plurals end in *-s*, such as *cats* and *dogs*, are similar.) The list of regular verbs is also open-ended. There are thousands, perhaps tens of thousands, of regular verbs in English (depending on how big a dictionary you consult), and new ones are being added to the language all the time. When *fax* came into common parlance a decade or so ago, no one had to inquire about its past-tense form; everyone knew it was *faxed*. Similarly, when other words enter the language such as *spam* (flood with E-mail), *snarf* (download a file), *mung* (damage something), *mosh* (dance in roughhouse fashion), and *Bork* (challenge a political nominee for partisan reasons), the past-tense forms do not need separate introductions: We all deduce that they are *spammed, snarfed, munged, moshed*, and *Borked*.

Even young children do it. In 1958 the psychologist Jean Berko Gleason tested four- to seven-year-old children with the following procedure, now known as the *wug*-test:



This is a wug.



Now there is another one.
There are two of them.
These are two ________.

The children could have refused to answer on the grounds that they had never heard of a wug and had never been told how to talk about more than one of them. Instead, Berko Gleason wrote, "Answers were willingly, and often insistently, given." Three-quarters of the preschoolers and 99 percent of the first-graders filled in the blank with *wugs*. Similarly, when shown a picture of a man who knows how to *rick* or *bing* or *gling* and did the same thing yesterday, most children said that he *ricked* or *binged* or *glinged*.

The children could not have heard their parents say *wugs* or *binged* before entering the lab, because these words had been coined especially for the experiment. Children therefore are not parrots who just play back what they hear. And the children could not have been previously rewarded by parents for uttering those forms, because the children did not know the words before entering the lab. Children therefore are not like pigeons in a Skinner box, who increase or decrease the frequency of responses in reaction to the contingencies of reinforcement. Noam Chomsky and Eric Lenneberg, pioneers of the modern study

of language and contemporaries of Berko Gleason in the Harvard-MIT community, pointed to children's ability to generalize constructions such as the regular past tense in support of their theory that language is actively acquired by a special rule-forming mechanism in the mind of the child.[15]

As it happens, all children are subjects in a version of Berko Gleason's experiment. Children often make up words or mangle them and are happy to put their new verbs in the past tense. Here are some examples:

spidered
lightninged
smunched
poonked
speeched
broomed
byed (went by)
eat lunched
cut-upped egg[16]

All children also make creative errors in their speech like these:

I buyed a fire dog for a grillion dollars.
Hey, Horton heared a Who.
My teacher holded the baby rabbits and we patted them.
Daddy, I stealed some of the people out of the boat.
Once upon a time a alligator was eating a dinosaur and the dinosaur was eating the alligator and the dinosaur was eaten by the alligator and the alligator goed kerplunk.[17]

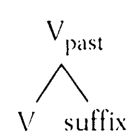
Such errors bring us to the second flavor of a verb in English: irregular. The past-tense form of an irregular verb is not simply the verb decorated with an *-ed* ending. For example, the past tense of *buy* is not *buyed*, but *bought*. Similarly, the past tense of *hear*, *hold*, *steal*, and *go* are *heard*, *held*, *stole*, and *went*.

Irregular verbs contrast with regular verbs in almost every way. Whereas regulars are orderly and predictable, irregulars are chaotic and idiosyncratic. The past tense of *sink* is *sank*, and the past tense of *ring* is *rang*. But the past tense of *cling* is not *clang*, but *clung*. The past tense of *think* is neither *thank* nor *thunk*, but *thought*. And the past tense of *blink* is neither *blank* nor *blunk* nor *blought*, but a regular form, *blinked*. The language maven Richard Lederer wrote a poem, "Tense Times with Verbs," that begins:

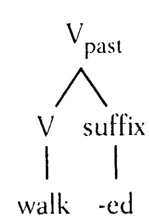
The verbs in English are a fright.
How can we learn to read and write?
Today we speak, but first we spoke;
Some faucets leak, but never loke.
Today we write, but first we wrote;
We bite our tongues, but never bote.
Each day I teach, for years I taught,
And preachers preach, but never praught.
This tale I tell; this tale I told;
I smell the flowers, but never smold.
If knights still slay, as once they slew,
Then do we play, as once we plew?
If I still do as once I did,
Then do cows moo, as they once mid?[18]

Also in contrast to the regulars, irregular verbs form a closed list. There are only about 150 to 180 irregular verbs in modern English (depending on how you count), and there have been no recent additions.[19] The youngest irregular is probably *snuck*, which sneaked into the language over a century ago and is still not accepted by purists.[20] And the freewheeling children in Berko Gleason's study were downright stodgy when it came to irregular forms: Only one out of eighty-six turned *bing* into *bang*, and one other turned *gling* into *glang*.[21]

These differences suggest a simple theory. Regular past-tense forms are predictable in sound and generated freely because they are products of a rule that lives in the minds of children and adults: "The past tense of a verb may be formed from the verb followed by the suffix *-ed*." The rule would look just like the rules of syntax in the toy grammar we played with earlier,



and would generate a similar inverted-tree-like structure:



Irregular verbs, in contrast, are unpredictable in form and restricted to a list because they are memorized and retrieved as individual words. An irregular form would look just like the lexical entry we saw when considering the name of the rose. It would be linked with the entry for the plain form of the same verb and labeled as its past tense:

hold ————	held
sound: *hōld*	sound: *hĕld*
meaning: ✍	meaning: ✍
part of speech: V	part of speech: V
	tense: past

Two mechanisms trying to do the same job would get in each other's way unless something adjudicated between them, and there is indeed a simple principle: If a word can provide its own past tense from memory, the rule is blocked; elsewhere (by default), the rule applies.[22] The first part explains why we adults don't say *holded* and *stealed*; our knowledge of *held* and *stole* blocks the rule that would have added *-ed*. The second part explains why both children and adults say *Borked* and *moshed* and *ricked* and *broomed*; as long as a verb does *not* have a form in memory, the rule may be applied. The ability of a rule to apply *elsewhere* or by default—that is, to any word that does not already have a specified form in memory—is the source of its power. A speaker who needs to express a past tense or plural is never left speechless, even when a search in memory comes up emptyhanded.

The theory that regular forms are generated by rule and irregular forms are retrieved by rote is pleasing not only because it explains the differences in productivity between the two patterns but also because it fits nicely into the larger picture of the design of language.

At first glance irregular verbs would seem to have no reason to live. Why should language have forms that are just cussed exceptions to a rule? What are they good for, besides giving children a way to make cute errors, providing material for humorous verse, and making life miserable for foreign language students? In Woody Allen's story "The Kugelmass Episode" a humanities professor in a midlife crisis finds a magic cabinet that projects him into any book he takes in with him. After a tempestuous affair with Madame Bovary, Kugelmass tries again with another novel, but this time the cabinet malfunctioned, and the professor "was projected into an old textbook, *Remedial Spanish*, and was running for his life over a barren, rocky terrain as the word *tener* ('to have')—a large and hairy irregular verb—raced after him on its spindly legs."[23]

But under the word-and-rule theory we need not suppose that evolution fitted us with a special gadget for irregularity. Irregular forms are just words. If our language faculty has a knack for memorizing words, it should have no inhibitions about memorizing past-tense forms at the same time. These are the verbs we call irregular, and they are a mere 180 additions to a mental lexicon that already numbers in the tens or hundreds of thousands. Irregular and regular forms therefore would be the inevitable outcome of two mental subsystems, words and rules, trying to do the same thing, namely, express an event or state that took place in the past.

Regular and irregular forms throw a spotlight on the advantages and disadvantages of words and rules, because everything else about them is the same: They both are one word long, and both convey the same meaning, past tense. The advantage of a rule is that a vast number of forms are generated by a compact mechanism. In English the savings are significant: The rules for *-ed*, *-s*, and *-ing* (the three regular forms of the verb) cut our mental storage needs to a quarter of what they would be if each form had to be stored separately. In other languages, such as Turkish, Bantu, and many Native American languages, there can be hundreds, thousands, or even millions of conjugated forms for every verb (for different combinations of tense, person, number, gender, mood, case, and so on), and the savings are indispensable. The rule also allows new words like *mosh*, rare words like *abase*, and abstract words like *abet* to be supplied with a past tense (*moshed*, *abased*, *abetted*), even if there were no previous opportunities for the speaker and hearer to have committed the form to memory. On the other hand, a rule is more powerful than needed for words we hear so often that retrieval from memory is easy. As we shall see, it is the most common verbs, such as *be*, *have*, *do*, *go*, and *say*, that turn out to be irregular in language after language.

Rules have another shortcoming that invites the word system to memorize irregulars. Recall that one of the nuisances plaguing John Wilkins as he designed his perfect language was that flesh-and-blood humans had to pronounce and understand the products of the rules. A sequence of sounds that encodes a concept precisely and efficiently may be unresolvable by the ear or unpronounceable by the tongue. So it is with the rule for the past tense in English. The delicate tongue-tap that graces the end of a regular form may escape a listener and be omitted when he reproduces it, resulting in a solecism such as *suppose to*, *use to*, or *cut and dry*, or in signs and inscriptions like these:

Broil Cod
Use Books

Whip Cream
Blacken redfish
Can Vegetables
Box sets
Handicap Facilities Available

In certain older expressions *-ed* was omitted so often that the expression eventually lost the *-ed* altogether, even among careful speakers and listeners. That's how we ended up with *ice cream* (originally *iced cream*), *sour cream*, *mince meat*, and *Damn Yankees*.[24] Irregular verbs, in contrast, tend to use vowel changes such as *ring–rang*, *strike–struck*, and *blow–blew*, which are as clear as a bell.

Similarly, the very obliviousness to the details of the verb that makes a rule so powerful (it applies across the board to all verbs, whether they are familiar sounding or not) can let it blindly jam a suffix onto the end of an inhospitable sound. The result can be an uneuphonious tongue-twister such as *edited* or *sixths*. Monstrosities like these are never found among the irregulars, which all have standard Anglo-Saxon word sounds such as *grew* and *strode* and *clung*, which please the ear and roll off the tongue.[25]

Language works by words and rules, each with strengths and weaknesses. Irregular and regular verbs are contrasting specimens of words and rules in action. These are the themes of this book, but with many twists to come. It would be too good to be true if we reached a major conclusion about the most complicated object in the known universe, the human brain, simply by seeing how children name pictures of little birds. The word-and-rule theory for regular and irregular verbs is an opening statement in the latest round of a debate on how the mind works that has raged for centuries. It has inspired two alternative theories that are equally ingenious but diametrically opposed, and intensive research showing what is right and wrong about each of them—perhaps resolving the debate for good. The theory has solved many puzzles about the English language, and has illuminated the ways that children learn to talk, the forces that make languages diverge and the forces that make them alike, the way that language is processed in the brain, and even the nature of our concepts about things and people. But to reach those conclusions we first must put regular and irregular verbs under a more powerful magnifying glass, where we will find some unexpected fingerprints.