as *this* and *that*, *here* and *there*, and *me* and *you*, the higher and farther-to-the-front vowels are found in the word that means "self" or "near the self," the lower and farther-to-the-back word means "other" or "far from the self." That is true not only in English but in many families of languages.[53]

Perhaps this ubiquitous vowel contrast is a case of sound symbolism. The linguist Roger Wescott has pointed out that high front vowels are pronounced with a constricted mouth cavity and the tongue close to the visible part of the vocal tract, whereas low back and central vowels are pronounced with a large mouth cavity and the tongue buried from view. That may call to mind the conceptual distinction between presentness and pastness. Pastness may remind people of a cavity or space, because a past event is separated by an interval of time from the present moment, and metaphorically speaking time equals space. It may also remind people of remoteness or distance, because metaphorically speaking long ago equals far away. Perhaps as Indo-European was developing, speakers vaguely felt that lower and farther back vowels fit better with the concept of an event separated in time from the present, and that higher and farther front vowels fit better with an event in the here and now.[54] Of course, the Indo-Europeans had to pick *some* vowel contrast if they were to mark tense with a vowel, and for all we know they could just have easily gone the other way. But the fact that the vowel contrast appears in many unrelated languages with similar roles, and was preserved and embellished in our own 5500-year game of Broken Telephone, hints that it might have some semantic resonance for human minds.

4

IN SINGLE COMBAT

Now that we know all about regular and irregular verbs, how well does the words-and-rules theory hold up? In some ways quite well; in other ways not so well. According to the theory, the ingredients of language are a list of memorized words, each an arbitrary pairing between a sound and a meaning, and a set of productive rules that assemble words into combinations. Regular and irregular forms exemplify the two ingredients: Regular forms are generated by rule, irregular forms are memorized by rote.

The dissection of language in chapter 2 showed that the organs that secrete regular forms are elegant combinatorial systems, just as the theory had led us to expect. But the excavation in chapter 3 showed that the depository for irregular verbs is not disorganized and inert, which is what the theory had led us to expect. The irregular verbs are shot through with patterns:

> *blow–blew, grow–grew, know–knew, throw–threw*
> *bind–bound, find–found, grind–ground, wind–wound*
> *drink–drank, shrink–shrank, sink–sank, stink–stank*
> *bear–bore, swear–swore, tear–tore, wear–wore*

This is not what we would expect if they were a laundry list of arbitrary items. Suppletive pairs such as *go–went* and *be–was* are the exception, whereas if the irregulars were truly acquired one by one they could just as easily have been the rule.

We have seen that many irregular patterns are fossils of extinct rules that lived in the heads of speakers long ago, but history can explain only part of the patterning. If the irregulars got all of their patterning from old rules and have degenerated steadily since the rules became extinct, today's patterns should be tattered versions of the ancient Indo-European strong classes, with a core of surviving verbs that fit the old patterns and a halo of distorted verbs that have drifted off in various directions. But many verbs have *joined* the strong classes, or have jumped from class to class, and today's irregular families cross-classify the original ones. *Ring–rang* originally was weak (with a past tense like *ringed*) and was attracted to the *ing–ang–ung* class by analogy to verbs like *sing*. The same happened to *dig–dug, stick–stuck, wear–wore, show–shown*, and many others. Some verbs that entered the language after the Old English period also were seduced into strong classes, such as *fling–flung* and *sling–slung*. Others switched from one strong class to another, such as *slay–slew* and *draw–drew* (originally *drough*). Still others found the *weak* irregular patterns appealing, such as *light–lit* and *creep–crept*.[1]

The words-and-rules theory would be off the hook if these attractions and conversions were a thing of the distant past. Perhaps the vowel-change rules did not die out completely with the Indo-European or Germanic tribes, but lingered in weakened form in parts of England for a few centuries before giving up the ghost. Unfortunately, some of the conversions are fairly recent. *Kneel–knelt, dive–dove, catch–caught*, and *quit–quit* became popular only in the nineteenth century; George Washington, for example, used *catched*, and Jane Austen used *quitted*. And as we saw, *snuck* came into English a century ago and is only now becoming standard.

The funny irregulars in the nonstandard dialects of English add to the stories. While some rural irregulars are quaint holdovers of old standard forms, such as *help–holp, climb–clim*, and *creep–crope*, many, if not most, are homegrown products of the creativity of local speakers: *bring–brang–brung, dive–div, chide–chode, snow–snew, climb–clomb, drag–drug, slide–slud, fling–flang*, and literally hundreds of others.[2]

As an experimental psychologist I have been trained not to believe anything unless it can be demonstrated in the laboratory on rats or sophomores. To my knowledge no one has yet studied irregular verbs in rats, but the linguists Joan Bybee and Carol Moder have studied them in sophomores and have shown that they are all too happy to generalize irregular patterns to novel verbs. They asked the students in a university linguistics course (not only sophomores, of course) to write down the past-tense forms of existing and made-up verbs

completing sentences such as *Sam likes to spling. Yesterday he* ______. Almost 80 percent of the subjects offered *splang* or *splung*. Even when given real words, some of the subjects were tempted by irregular patterns and came up with creative forms such as *dig–dag, sting–stang, slink–slank, streak–strack, skid–skud*, and *clip–clap*. Bybee and Moder may have duplicated in the laboratory the process that gave us forms like *fling–flung* and *drag–drug* in the history of English.[3]

But maybe not. University students treat anything that looks like a test as a test, and perhaps they saw the task as a challenge to their ingenuity in coming up with creative forms. We know that outside the lab, people ruminate over irregular patterns and sometimes deduce ways in which they ought to be generalized. Here is Leo Rosten and Leonard Ross's character Hyman Kaplan, a Jewish immigrant learning English at night school:

> "I tought de pest time 'bite' should be—'bote.'"
>
> Miss Mitnick gave a little gasp.
>
> "'Bote'?" Mr. Parkhill asked in amazement. "'Bote'?"
>
> "'Bote'!" said Mr. Kaplan.
>
> Mr. Parkhill shook his head. "I don't see your point."
>
> "Vell," sighed Mr. Kaplan, with a modest shrug, "if is 'write, wrote, written' so vy isn't 'bite, bote, bitten'?"
>
> Psychic cymbals crashed in Mr. Parkhill's ears.
>
> "There is not such a word 'bote,'" protested Miss Mitnick, who took this all as a personal affront. Her voice was small, but desperate.
>
> "Not-soch-a-void!" Mr. Kaplan repeated ironically. "Mine dear Mitnick, don' *I* know is not soch a void? Did I say *is* soch a void? All I'm eskink is, isn't logical *should be* soch a void?"[4]

The students in the *spling* experiment also may have thought to themselves, "Isn't logical *should be* soch a void?," and may have concocted forms they never would have used when speaking naturally.

A related worry is that people make note of irregular forms, especially in other people's dialects, and consciously use them in wit and wordplay, as we discovered in the previous chapter and can see in these old jokes:

> A woman gets into a taxi in Boston's Logan airport and asks the driver, "Can you take me someplace where I can get scrod?" He says, "Gee, that's the first time I've heard it in the pluperfect subjunctive."

A friend of mine came across some cut flowers that were so spectacularly red she thought they must be fake. "These are amazing," she said. "Are they dyed?" The florist shook his head. "No, no, not at all," he said. "Just put 'em in water and they'll be fine."[5]

A man was on trial for pulling a woman down the street by her hair. The judge asked the arresting officer, "Was she drugged?" The policeman replied, "Yes, sir, a whole block."[6]

Verbal humor depends on the audience's recognizing that an odd word *is* odd; if the words were unexceptional, there would be no joke. So just because some wiseguy finds a novel past-tense form logical or amusing, it does not mean that the form is a natural product of the system he uses in everyday speech. Instead the form may be a product of one's intellectual faculties reflecting back onto one's language, an ability called metalinguistic awareness.

But one kind of generalization cannot be a product of conscious cogitation, namely, the errors preschool children make in their spontaneous speech:

It was neat—you should have sawn it!
Doggie bat me [bit].
The cheerios got aten by the Marky.
I know how to do that. I truck myself [tricked].
He could have brang his socks and shoes down quick.
And they swang into a roller coaster and we went with their cars and they were sliding and they did a leap.
Elsa could have been shotten by the hunter, right?
So I took his coat and I shuck it [shook].
This is the best place I ever sot [sat].
I bate Paul up [beat].
You mean just a little bitty bit is dranken?[7]

The psychologist Fei Xu and I combed through transcripts of the speech of nine children in an electronic archive and pulled out all the past tense and participle forms, 20,000 in all.[8] We found numerous irregular forms that are not standard English:

beat–bate	*crush–crooshed*	*fling–flang*	*say–set*
bite–bet	*fight–fooed*	*jump–janged*	*sleep–slep*
bring–brid	*fit–feet*	*lift–left*	*swing–swang*

Children don't make these errors very often—only in two tenths of 1 percent of the opportunities—but eight of the nine children made at least one while the tape was running, and we know that children continue to make them well into the school-age years. I have a drawing from a seven-year-old girl with the caption *Win I Wit to Git a crismis chre and it wus Sowing so i braing my umbrella*[9] and I remember that when I was twelve I had a persistent urge to use *raught* as a past tense of *reach*, on the analogy of *teach–taught*.

~

The irregular patterns refuse to die. Irregular verbs are supposed to be a list of arbitrary words memorized by rote, just like *duck* and *walk*, with only a trace of patterning left behind by long-defunct rules. Instead, people extract the patterns and extend them to new words, just as they do with the *regular* pattern in errors like *breaked*, in neologisms like *moshed*, and in the *wug*-test. The distinction between regular and irregular inflection, and therefore between words and rules, is not so clear anymore. Either the irregular patterns are generated by rules, just like the regular pattern, or linguistic productivity does not depend on rules in the first place but can arise from words via some ability to associate the patterns in known words with the patterns in new ones.

Both alternatives have been developed into famous, full-blown theories of the English past-tense system. Their clash is one of the most vigorous controversies in the modern study of the mind, echoing through psychology, linguistics, philosophy, computer science, and neuroscience. According to the theory of *generative phonology* developed by Noam Chomsky and Morris Halle, rules rule.[10] Every drop of patterning in past-tense forms, regular or irregular, is squeezed out into rules, and only the compressed, desiccated residue is stored in the mental lexicon. According to the theory of *parallel distributed processing* or *connectionism* developed by David Rumelhart and James McClelland, there are no rules: People store associations between the sounds of stems and the sounds of past-tense forms, and generalize the associations to new words if they are similar to old words.[11] Both theories invoke a single kind of mental computation to explain how people generate regular and irregular forms, but for generative phonology it's rules all the way down, whereas for connectionism it's memory all the way up.

A clash over irregular verbs may call to mind the remark that academic debates are heated because so little is at stake. But in this case something is at stake. The past-tense debate is the latest battle in a centuries-old disagreement over two very different ways of understanding the mind:

When a man reasons, he does nothing else but conceive a sum total from addition of parcels, or conceive a remainder from subtraction of one sum from another; which, if it be done by words, is conceiving of the consequence of the names of all the parts to the name of the whole, or from the names of the whole and one part to the name of the other part. . . . For REASON is nothing but *reckoning*.[12]

In this passage from *Leviathan*, written in 1651, Hobbes uses *reckoning* in the original sense of counting, calculating, or computing. For example, suppose the definition of "man" is "rational animal." Then if we are told that something is "rational" and an "animal" (names of parts) we can deduce it is a "man" (name of whole), and if we are told that something is a "man" (name of whole) we can deduce that it is a rational and that it is "rational" (name of one part) and "animal" (name of the other part). These steps could be laid out as mechanical instructions to recognize and copy words, a kind of symbol, and therefore could be "reckoned" or computed by someone who has no idea what the concepts "rational" and "animal" even mean. If the symbols are patterns in the brain rather than words on a page, and the patterns trigger other patterns because of the way the brain is wired, then we have a theory of thinking.

Among the people influenced by Hobbes was Leibniz, who was inspired as well by John Wilkins and other designers of artificial languages discussed in chapter 1. Leibniz took Hobbes literally when he said that reason is nothing but reckoning. He devoted much of his life to inventing a scheme that would perfect the computations underlying thought, turning arguments into calculations and making fallacies as obvious as errors in arithmetic. "Once this has been done," he wrote, "if ever further controversies should arise, there should be no more reason for disputes between two philosophers than between two calculators. All that will be necessary is that, pen in hand, they sit down together at a table and say to each other . . . 'let us calculate.'"[13] In one version of Leibniz's scheme, "man" is assigned the number 6, "animal" is assigned 2, and "rational" is assigned 3. Since $2 \times 3 = 6$, a rational animal must be a man; since $6 \div 3 = 2$, a man is not just any old rational being but specifically a rational animal. If the number for "monkey" is 10, one may calculate that monkeys are not men or vice versa, and that monkeys, while animals, are not rational.

The idea that intelligence arises from the manipulation of symbols by rules is a major doctrine of the school of thought called rationalism, generally associated with Leibniz and Descartes. When the symbols stand for words and the rules arrange them into phrases and sentences, we have grammar, the subject

of Cartesian linguistics, which later inspired Humboldt and then Chomsky. When the symbols stand for concepts and the rules string them into chains of inference, we have logic, which became the basis for digital computers, the artificial intelligence systems that run on them, and many models of human cognition.[15]

But symbol manipulation is not the only way the mind might work:

There appear to be only three principles of connection among ideas, namely, *resemblance*, *contiguity* in time or place, and *cause or effect*. Experience shows us a number of uniform effects, resulting from certain objects. When a new object, endowed with similar sensible qualities, is produced, we expect similar powers and forces, and look for a like effect. From a body of like color and consistence with bread we expect like nourishment and support.[16]

In this passage from his 1748 *Enquiry Concerning Human Understanding*, David Hume summarizes the theory of *associationism*, a major tenet of the school of thought called empiricism. The mind connects things that are experienced together or that look alike (Hume later eliminated cause and effect as a separate principle) and generalizes to new objects according to their resemblance to known ones. Just as the rationalists were obsessed by combinatorial grammar, the associationists were obsessed by memorized words. In his 1689 *Essay Concerning Human Understanding*, John Locke pointed to the arbitrary connection between words and things as the quintessential example of how the mind forms associations by contiguity in time: We learn *dog* when Mother says "dog" in the presence of a dog. Replace Locke and Hume's "ideas" or "sensible qualities" with "stimuli" and "responses," and you get the behaviorism of Ivan Pavlov, John B. Watson, and B. F. Skinner. Replace the ideas with "neurons" and the associations with "connections," and you get the connectionism of David Rumelhart and James McClelland.

The great debate between rationalism and empiricism is familiar to everyone who has taken a course in philosophy, psychology, or the history of ideas. It embraces such issues as whether the mind is packed with innate structure or is a blank slate on which the environment writes, and whether knowledge comes from making deductions using theories or gathering data from observation. But the issue that concerns us here is the nature of our mental machinery, in particular, whether intelligence arises from the manipulation of symbols or from associations between sensory qualities. And the competing theories of the English past tense provide us with an unusual opportunity.

It's been three hundred and fifty years since *Leviathan*, and scholars are still debating rationalism and empiricism. Both sides appeal to theoretical coherence, intuitive plausibility, political ramifications, and harmony with the morals of modern science, but these are pretty squishy criteria, and the debate rages on. We need a concrete, richly studied instance of human psychology in which the two grand theories can go head-to-head in explaining the same facts.

In ancient warfare, an army sometimes would send its mightiest warrior to face his counterpart from the opposing army in single combat. The outcome would embolden one side in the actual battle that followed, or might pre-empt it altogether, sparing unnecessary bloodshed. The most familiar example is the biblical story of David and Goliath, and in *The Right Stuff* Tom Wolfe suggested that a contemporary example may be found in the space race of the 1960s, in which the Mercury astronauts were treated like single combat warriors against the cosmonauts of the Soviet Union. Scientific debates sometimes work like single combat, not because the combat metaphor is particularly apt or appealing, but because it is easier to compare two great big theories when each is vested in a highly specific hypothesis and the hypotheses compete on the same ground.

The English past tense is the perfect site. The phenomenon is circumscribed and therefore tractable to study. Its history, acquisition, and patterns of use have been abundantly documented. It has obvious rulelike and memory-like features that serve as hurdles that any theory must clear. And it has brought out the best in the ingenuity of contemporary theoreticians, with each side devising a clever, elegant, detailed, and surprising model. The past tense is the only case I know in which two great systems of Western thought may be tested and compared on a single rich set of data, just like ordinary scientific hypotheses.

~

What are the facts to be explained? Irregular verbs defy the suggestion that they are memorized by rote because they show three kinds of patterning.

First, irregular past-tense forms are similar in sound to their base forms. For example, *drink* and *drank* share *d*, *r*, a vowel, *n*, and *k*; the only difference is that *drank* has the vowel *a* where *drink* has the vowel *i*. Similarly similar are *swear* and *swore*, *sleep* and *slept*, *freeze* and *froze*. In fact, all the irregulars except *go–went* and *be–was* share material with their stems. It didn't have to be that way. One can imagine a hypothetical language in which most of the verbs are like *go–went*, with nothing in common between stem and past, each stem and each past stuffed into its own memory slot. We need an explanation of why English does not look like that. Let's call this pattern *stem-past similarity*.

Second, a few kinds of change from a stem to its past are seen over and over among the 164 irregular verbs. The *ĭ-ă-ŭ* pattern in *drink–drank–drunk*, for example, is found, with variations, in *sing–sang–sung*, *sit–sat–sat*, *begin–began–begun*, *shrink–shrank–shrunk*, and twenty other verbs. Similarly, we have *freeze–froze*, *speak–spoke*, and *steal–stole*; *bleed–bled*, *breed–bred*, and *feed–fed*; *teach–taught*, *fight–fought*, and *bring–brought*. One can imagine a language in which every verb picked its own substitution of vowels and consonants from among the thousands that are logically possible. But generations of learners have passed down an English language that is very different from that possibility. Let's call this pattern, in which the change from stem to past in one verb is similar to the change from stem to past in another verb, *change-change similarity*.

Third, the verbs undergoing a given irregular change are far more similar than they have to be. If you are a verb and want to undergo the *ĭ-ă-ŭ* pattern, all you really need is an *ĭ*. But the verbs that do follow the pattern (*drink*, *spring*, *shrink*, and so on) have much more in common; most begin with a consonant cluster like *st-*, *str-*, *dr-*, *sl-*, or *cl-*, and most end in *-ng* or *-nk*. Similarly, the verbs whose pasts end in *-ew* (*blow*, *grow*, *throw*, *slay*, *draw*, and *fly*) tend to begin with a consonant cluster and end with a vowel. Verbs of a feather change together, and not just in sightings by word watchers. People extend old patterns to new verbs, as in *bring–brang*, *fight–fit*, and *spling–splung*, only when the new verb is highly similar to old ones in memory. We need an explanation of why the human mind is so impressed by similarity in sound; let's call this pattern *stem-stem similarity*.

A theory with rules for *irregular* verbs, as well as regular ones, could explain all three kinds of patterning. Imagine a rule that said, "If a verb has the sound *consonant-consonant-ĭ-ng*, change *ĭ* to *ŭ*." Notice that the rule doesn't spell out the past-tense form letter by letter; it just says, "Change the vowel." The rest of the input—the consonants before and after the vowel—come through in the output untouched. We have an explanation for stem-past similarity.

Now suppose that the mind prefers simple grammars, with a few rules, to complex grammars, with many rules. If there are fewer rules than verbs, many verbs will have to share a rule, such as "Change *ĭ* to *ŭ*." We have an explanation for change-change similarity.

Finally, notice that the rule has a condition on it: apply only to verbs that have two consonants before the vowel and an *ng* sound after it. The condition is a gatekeeper that allows in verbs that are similar to the *cling–clung* family and filters out those that merely contain ĭ. This could explain stem-stem similarity.

Now, any drudge can go over the list of irregular verbs in the preceding chapter and write down a set of tedious rules for them: "If a verb begins in *s* and ends in *ee*, change the *ee* to *aw*," and so on. But that would be no improvement over the original list of verbs. A theory invoking rules must be more than a summary of the patterns among the verbs. It should be a *psychological* theory: a hypothesis of the format in which children acquire words, and an explanation of why verbs have the kinds of patterns they do. The trick is to find a compact set of rules that captures the generalizations the mind likes to make.

By far the most ambitious theory of this kind comes from Noam Chomsky and Morris Halle's 1968 magnum opus *The Sound Pattern of English*, later refined by Halle and the linguist K. P. Mohanan.[17] Their rules for irregulars are part of a larger set of rules that capture the sound pattern of a language (its accent). Clearly the speakers of a language know more than the list of words that happen to be in the language at a given moment. For example, English speakers intuit that *blicket*, *dax*, and *fep* are not English words but could be, whereas *ftip*, *rtut*, and *nganga* are not English words and could not be (though speakers of other languages might recognize them as possible words in their languages). English speakers also know that when *divine* is joined to *-ity* to create *divinity*, the *i* vowel in *-in-* changes from ī to ĭ, and that when *Canada* is joined to *-ian* to become *Canadian*, the final *-a* vanishes, the stress shifts from the first syllable *Ca* to the second syllable *na*, and the vowel in that syllable changes to ā. Chomsky, Halle, and Mohanan accounted for the patterns of thousands of English words with just a few dozen phonological rules, each assigned to one or more of the boxes in the diagram on page 23: lexicon, morphology, syntax. Their theory comes from a field called *generative phonology*, a division of generative linguistics, the approach to language founded by Chomsky.

By positing rules that replace consonants and vowels (phonemes) in the irregular verbs, Chomsky and Halle enjoy the advantage of rules in general: accounting for patterns among verbs and changes, and for speakers' ability to generalize them. Amazingly, Chomsky, Halle, and Mohanan handled most of the dizzying patterns among the 165-odd irregular verbs with only *three* rules. Virtually all their other rules are needed to explain the sound pattern of English in general.

Chomsky, Halle, and Mohanan roundly reject the words-and-rules dichotomy. Verbs sit on "a continuum of productivity and generality that extends from affixation of the *-ed* suffix in *decide–decided* to total suppletion in *go–went*," with fam-

ilies like *sing–sang, ring–rang*, and *bind–bound, wind–wound* in between.[18] At one end of the continuum are the regular verbs, which are handled by a general rule that says nothing about the words it can apply to. At the other end of the continuum are suppletive verbs such as *go* and *went*, which are simply listed as pairs. In between are the other irregulars, which are handled by a smaller set of rules, each tagged to apply to certain verbs. By stipulating which verb may be touched by which rule, the theorists circumvented the problem of crafting the rules to single out verbs by their sounds—no small matter, given that *shrink* has the past tense *shrank*, *sling* has the past tense *slung*, *bring* has the past tense *brought*, and *blink* has the regular past tense *blinked*.

Another rein on rules keeps each one in a *stratum* (a component or subcomponent) so that it does not run wild and apply where it shouldn't. To generate *keep–kept*, for example, Chomsky, Halle, and Mohanan invoked a rule that shortens a long vowel (changing ē to ĕ) when it occurs before a consonant cluster, such as *pt*. But that rule cannot be allowed to apply across the board or it would turn *seeped* into *sept*, *wiped* into *wipped*, and so on. So Halle and Mohanan proposed that the *-t* and *-d* found in weak irregular verbs like *kept* is not the same as the *-ed* found in regular verbs, despite their similar pronunciations. Whereas the regular *-ed* is attached in the morphology box on page 23, the *-t* or *-d* in the irregular verbs is attached in the lexicon box, which is also where the shortening rule is confined. This may seem like cheating, but there are independent grounds for it. Other semiregular sort-of-rules that have nothing to do with the past tense, like the ones creating *serene–serenity* and *volcano–volcanic*, need the shortening rule too. That supports the idea that several rules are sequestered together in their own little community.

One move that allowed Chomsky, Halle, and Mohanan to get away with so few rules was factoring apart each complex change into several simple ones and allowing the simple ones to be assigned in different combinations to different verbs. For example, *tell–told* is produced by at least two rules, one that changes the vowel, the other that adds the *-d*. The rule that changes the vowel is also put to work in *swear–swore*, and the rule that adds *-d* is also put to work in *flee–fled*. Looking at the crisscrossing patterns in *tell–told, swear–swore, flee–fled, bend–bent, burn–burnt, deal–dealt, breed–bred*, and *hit–hit*, one can readily see that sharing minirules that add *ts* and *ds*, that delete extra ones, and that fiddle with vowels is far more economical than building a special rule for each family of rhyming verbs.

A monumental contribution of generative phonology was to slice rules even more finely so that they apply not to vowels and consonants but to the *compo-*

nents of vowels and consonants, called *features*. The idea goes back to Roman Jakobson and is a striking, universal trait of human language. In chapter 2 we noticed that the three pronunciations of *-ed* in *walked*, *jogged*, and *patted* are determined by the final consonant of the stem:

ɨd if the verb ends in *t* or *d*
t if the verb ends in *p, k, f, s, sh, ch*, or unvoiced *th*
d if the verb ends in a vowel or in *l, r, m, n, b, g, v, z, j, zh* or voiced *th*

Each list is not just any old collection of consonants. The two consonants in the first line, *t* and *d*, are pronounced with the same parts of the mouth (tongue tip against gum ridge) and in the same way (stopping the flow of air and then releasing it). These features—"place = tongue-tip," and "manner = stopping"—are also found in the *-ed* suffix. A similar pattern of sharing also is found among *s, z*, and the plural suffix, all of which are sibilants. What we never see is a rule such as "Add *-og* if the word ends in *z, r*, or *k*," and other ragtag assemblies of phonemes.

All this can be captured if we have rules apply to *features* rather than to *phonemes*. One rule can state, "At the end of a word, insert *ɨ* to separate adjacent consonants that have similar features for place and manner of articulation"—no listing of *t, d, s, z, -ed*, or *-s* is necessary. Similarly, it's obtuse to have a rule that puts a *t* after consonants in the list *p, k, f, s, sh, ch*, and *th*—all these consonants obstruct the air stream (they are *obstruent*) and all are unvoiced. The rule can simply say, "At the end of a syllable, copy the voicing feature from one obstruent consonant to the next." All the unvoiced consonants like *p* and *k* will automatically get a *t*, while all the voiced consonants like *b* and *g* automatically get a *d*. Moreover, this rule also generates the *s* and *z* variants of the plural suffix.

Simple rules that inspect, flip, or excise features are ubiquitous in the world's languages. They not only are more economical than rules that hack away at arbitrary lists of consonants and vowels, but also allow speakers to generalize. The sound *ch* is not normally found in English words, but any English speaker who labors to pronounce the celebrated composer's name as *Bach* knows that if there were a verb *to out-Bach*, as in *Handel out-Bached Bach*, the past tense would be pronounced *Bacht*, not *Bachd* or *Bachɨd*. A rule that specifies *t* after "unvoiced consonants" automatically embraces the unvoiced *ch* and tells a speaker what to do, even if the speaker had never learned that *ch* belongs on the list.

This economy and power also accrues to rules that fiddle with *vowels* if the vowels are dissolved into features such as these:

Is the tongue hump at the front or the back of the mouth?
Is the tongue hump high or low in the mouth?
Are the lips rounded or not?
Is the vowel long or short?
Is the vowel tense (tongue root scrunched forward) or lax?

Take *sing–sang* and *sit–sat*. The *ĭ* is not replaced by some random vowel, such as the one in *say* or *boat* or *shoe*; it is replaced by *ă*, which is identical to *ĭ* except for tongue height: *ĭ* is front, unround, short, lax, and high; *ă* is front, unround, short, lax, and low. A rule that simply said, "For the following irregular verbs, lower the vowel" would have to tinker with only one feature, not five; it would explain why *ĭ* was replaced by a similar vowel, not just any old vowel; and it would work right out of the box to yield *eat–ate* and *choose–chose*, which also lower a vowel. This simple rule, Lowering Ablaut, is one of the three irregular rules that Halle and Mohanan's theory gets away with. The other two are Backing Ablaut, which replaces mid front *e* in *bear* with mid back *o* in *bore*, and Shortening Ablaut, which replaces the long vowels in *flee* and *shoot* with their short counterparts in *fled* and *shot*.

If you have been skimming this last paragraph silently, everything should look shipshape—the vowels in *flee–fled* and *shoot–shot* are literally spelled as long and short versions of the same thing: *ee* versus *e*, *oo* versus *o*. But if you have been pantomiming the sounds with your mouth and listening to them with your mind's ear, or better yet, pronouncing them out loud, you should be thinking: *Now wait a minute!!* As we saw in the last chapter, "long" and "short" have been misnomers in English at least since the Great Vowel Shift in the fifteenth century, when people scrambled the pronunciations of vowels. The sound in *flee* is *not* a drawn-out version of the sound in *fled*, nor is *shoot* just a lengthy *shot*. In pronouncing *ee* (*ē*) the tongue is higher than it is when pronouncing *ĕ* and more tensed up, and the vowel glides up to a little *y* at the end, making it a diphthong (a succession of two vowels pronounced as if they were one). Likewise, in pronouncing *oo* (*ū*), the tongue is tenser and higher than when pronouncing *ŏ*, the lips are rounded, and there is a little *w* sound at the end. Those vowels are not particularly similar, and a rule capable of replacing one with another is capable of doing almost anything. To say that a Shorten-the-Vowel rule simplifies the hairy irregular verbs sounds like a hoax.

Chomsky and Halle realized this of course, and their solution is the most radical claim of the theory. When it comes to syntax, Chomsky is famous for proposing that beneath every sentence in the mind of a speaker is an invisible, inaudible *deep structure*, the interface to the mental lexicon. The deep structure is converted by transformational rules into a "surface structure" that corresponds more closely to what is pronounced and heard. The rationale is that certain constructions, if they were listed in the mind as surface structures, would have to be multiplied out in thousands of redundant variations that would have to have been learned one by one, whereas if the constructions were listed as deep structures, they would be simple, few in number, and economically learned (see *The Language Instinct*, chapter 4, 120–124). Less well known is that Chomsky and Halle made a similar proposal for the sounds of words. Each word has a deep structure—in jargon, an *underlying form*—that may not sound like the way it is pronounced; indeed, it may be unpronounceable. Phonological rules then convert it to the surface form that is articulated and heard.

In the case of the so-called long vowels in English, Chomsky and Halle proposed that the underlying forms really *are* long versions of the short vowels. That is, in the mental lexicon the vowels in the following pairs are *identical* in every respect except how long it would take to pronounce them, with the long vowels taking about twice as long:

Short Vowels	Their Long Counterparts
din	*diviin (divine)*
den	*sereen (serene)*
pat	*saan (sane)*
fund	*profuund (profound)*
shot	*shoot (shoot)*
bomb	*coon (cone)*

In their actual pronunciations these pairs do differ in how long it takes to say the vowel, but they differ in many other ways besides, so why assume that the mind lists only the difference in length? The reason, according to Chomsky and Halle, is that the other differences are redundant and predictable, hence unnecessary to list. No pair of English words differs *only* by vowel length. Long vowels are also tense and diphthongs (they glide to a different vowel at the end); short vowels are lax and not diphthongs. A theory that gave the mind the ability to store every nuance of an English vowel—length, tenseness, tongue position, lip rounding, diphthongs, and so on—would falsely predict that English could contain long lax vowels, lax diphthongs, short tense vowels, and so on; and it would allow the language to contain short and long versions of otherwise identical vowels—all counter to fact. In a better theory, the vowel inventory would contain only the number of vowels necessary to distinguish words (for example, to distinguish *bit* from *bet*), and other rules would fill the vowel out into a full set of stage directions for the mouth and throat. The best way to do this is to list certain pairs of vowels as differing only by being long or short, and to have the long versions trigger obligatory rules that flesh out the rest of their pronunciation.

Chomsky and Halle therefore proposed that English has a rule of Long Vowel Tensing, which tenses all long vowels, and a rule of Diphthongization, which adds the little *y*s and *w*s that give us the two-part vowels in *lake* (leh-eek), *glide* (gla-eed), *need* (nee-y'd), *loud* (la-ood), and *road* (ro-ood). All this is more or less unexceptionable—*something* in the mind of an English speaker enforces a correlation among length, tenseness, and being a diphthong, and the predictable details of pronunciation need not be stored in individual lexical entries.

The theory that a word is stored as an abstract, not-directly-pronounceable deep structure has another advantage, pointed out by the linguist Aditi Lahiri and the psychologist William Marslen-Wilson.[19] Consider the seemingly simpler idea that memory holds the actual pronunciation of a word. The problem is: *which* actual pronunciation of the word? The word *hand*, for example, might be pronounced *h-a-n-d* when we enunciate it carefully and distinctly, but in natural conversation it comes out quite different. The *nd* is pronounced as *nj* in *hand you*, as *m* in *hand me*, and as the *ng* sound in *hand care*. (That is one of the reasons why computer speech recognition systems, though pretty good at recognizing words in isolation, are still poor at recognizing words in connected speech.) But if, as Chomsky and Halle proposed, the dictionary entries for words are schematic—so that the last segments of *hand* are listed, say, as "nasal" and "dental" rather than as *n* and *d*, and they are fleshed out into full consonants by rules that work in different ways in different contexts—then a single representation could embrace the *hand* that appears in *hand*, *hand me*, *hand you*, and *hand care*.

But Chomsky and Halle went much further than merely claiming that words have a form in memory that is not identical to their pronunciations. They proposed that the underlying form of a word can be *wildly different* from its pronunciation. In particular, they proposed a complicated rule of Vowel Shift that raises or lowers the long vowels, reenacting in the minds of modern English speakers the Great Vowel Shift of the fifteenth century. In other words, they claim that ontogeny recapitulates phylogeny, and that the deep structures of words in our mental dictionaries correspond to the way Chaucer would have pronounced them (even though Chaucer, if he traveled through time to our century, would sound like a German to our ears). According to the Chomsky-Halle theory, the mental representations of words in different centuries over the past millennium, and in all the modern dialects of English, are the same; English has changed primarily by adding phonological rules. And English spelling, which did not track the Great Vowel Shift or other changes in pronunciation as the dialects evolved, captures our underlying mental representations of words.

Chomsky and Halle pursue the implications. Everyone agrees that a good spelling system ought to be stable across time and space. We should be able to read the writings of our great-great-grandparents, and of people on the other side of the Atlantic, even if they pronounce words differently from the way we do. Also, a spelling system ought to encode only the information necessary to identify the content of a word, not the trajectories of lip pursing and tongue flicking that can be predicted from the content and that people automatically execute as they talk. By these criteria, Chomsky and Halle concluded, English spelling is not only exonerated of the charge that it is an illogical, sadistic mess, but "comes remarkably close to being an optimal orthographic system."[20] Optimal for us, optimal for other modern English dialects, and optimal for all the recorded dialects of the past several centuries![21] Hear that, all you orthographically challenged, spell-checker-dependent, solecism-prone students and writers? Forget *cough* and *rough* and *dough* and *plough*, and *ghoti* spelling *fish*, and George Bernard Shaw's campaign to reform English spelling, and all the other complaints about crazy English:

> A *moth* is not a *moth* in *mother*,
> Nor *both* in *bother*, *broth* in *brother*,
> And *here* is not a match for *there*,
> Nor *dear* and *fear* for *bear* and *pear*.
> And then there's *dose* and *rose* and *lose*
> Just look them up—and *goose* and *choose*,
> And *cork* and *work* and *card* and *ward*,
> And *font* and *front* and *word* and *sword*,
> And *do* and *go* and *thwart* and *cart*—
> Come, come, I've hardly made a start!
>
> A dreadful language? Man alive!
> I'd mastered it when I was five.
> And yet to write it, the more I tried,
> I hadn't learned at fifty-five.[22]

What led Chomsky and Halle to this shocking conclusion? It was the drive to extirpate any trace of needless redundancy and complexity in their grammar for English sound patterns. Now we really do have a clean Shortening Ablaut rule for *breed–bred*, *flee–fled*, *shoot–shot*, and *lose–lost*. The underlying form of *breed* has a double-length version of the *e* in *bred*, so the shortening rule creates *bred* in a single step. (The formerly inconvenient fact that *breed* itself is not double-length *bred* is now handled by Vowel Shift, which makes it *briid*, followed by Tensing, which makes it *breed*, followed by Diphthongization.) Now if it was only a handful of irregular verbs that benefited from the Vowel Shift rule, the savings would be paltry compared to simply stipulating that "$\bar{e}$ changes to $\breve{e}$," and so on. The savings begin to mount, however, when we look at *other* rules that can be simplified in exactly the same way, that is, by applying to the deep, pre-Shifted versions of vowels. With Vowel Shift available to handle the details of the long vowel, each of the following processes can be captured as a simple change-the-length rule:

Trisyllabic Shortening:	*divine–divinity*	*serene–serenity*	*sane–sanity*
Cluster Shortening:	*crucify-crucifixion*	*intervene–intervention*	
-ic Shortening:	*satire–satiric*	*kinesis–kinetic*	*volcano–volcanic*
CiV Lengthening:	*study-studious*	*manager–managerial*	*Canada–Canadian*

Once you have the freedom to equip people with abstract underlying forms for their words, the irregulars get simpler and simpler. *Run–ran–run* can be handled by the rules for *sing–sang–sung*, *drink–drank–drunk*, and so on—if you suppose that the underlying form of *to run* is really *to rin*, and that Backing Ablaut and other rules apply to the *stem*, not just the participle, to make it surface as *run*. Likewise, the past-tense forms of *come*, *give*, *slay*, and *catch* are better behaved if their underlying stems are *kēm*, *gēv*, *slē*, and *kĕch*.

Most creatively of all, Chomsky, Halle, and Mohanan proposed that the *ch* of *Bach* is a covert English phoneme that lives underground in the lexical entries for *buy* and *fight*—namely, *bēch* and *fēcht*—and in the half-baked past-tense forms for *seek* and *teach*. Of course the *chs* must be assassinated before they see the light of day, but not until they have triggered a rule that makes the past-tense form come out right. *Bēch* gets a *-t*, changes to *bōcht* by the Lowering and Backing Ablaut rules, at which point the *cht* triggers Cluster Shortening to yield *bŏcht* before *ch* makes the ultimate sacrifice, resulting in the form we spell *bought*.

~

What are we to make of this bold theory? As I mentioned at the start, a theory that posits rules for irregulars can account for the similarities between stems and their past-tense forms, such as why *swing* and *swung* are 80 percent the same: The rule targets a vowel for change, and leaves the rest of the verb alone. The Chomsky-Halle-Mohanan theory pushes the performance of rules to new heights, because their rules target only certain *features* of a vowel for change (such as tongue height or vowel length) and leave the rest of the vowel alone, too. Similarly, a theory positing irregular rules can account for the similarities in changes undergone by different verbs, for example, why the *ĭ-ă* pattern in *sing–sang* is also found in *drink–drank* and *sit–sat*: A few rules are shared by many verbs. Here the Chomsky-Halle-Mohanan theory succeeds with a vengeance, forcing almost 165 verbs to share only three rules.

Any theory that can tame the quintessentially unruly English irregular past-tense system with only three rules, each delicately adjusting a single feature, is undeniably brilliant. But is it true? Not necessarily. One problem comes from the assumption that every scintilla of patterning in the verb system needs an explanation in terms of the psychology of speakers, in particular that the patterns are distilled out into rules in the mind. Chomsky, Halle, and Mohanan's rule-by-rule derivations often recapitulate the history of a past-tense form in English over the centuries—deliberately—and that brings to mind an alternative explanation used throughout chapter 3: that the patterns are fossils of rules that died long ago. The surviving past-tense forms, semilawful though they are, could simply be memorized by today's generation without any help from the rules.

The defunct-rule explanation has an advantage over the Chomsky-Halle-Mohanan theory. Children don't hear underlying forms, and they are not provided with lessons about the rules that turn them into audible surface forms. They hear only the surface forms. If the rules and underlying forms are to play some role in mental life, children must infer the cascade of rules that generated the surface form, run it in reverse, and extract the underlying form. And the suggestion that English-speaking children hear *run* and infer *rin* or hear *fight* and infer the German-sounding *fēcht* is, frankly, beyond belief.

First, why would the child bother if the rules are there only to generate the surface form, and the child already *has* the surface form? (It's different with sentences, where the child needs rules to generate an infinite number of new ones; with word roots, there are only a finite number to learn.) And even if the child wanted to ferret out rules and underlying forms, how could they ever *find* the right ones if the crucial clues—the ones linguists themselves use to discover the rules—are found in pairs of words the children will learn only in adulthood if ever, such as *serene* and *serenity*, *manager* and *managerial*, *kinesis* and *kinetic*? At one point Chomsky and Halle concede the problem and say that their grammar is only what children *would* construct if, hypothetically, they could hear the entire vocabulary in one sitting before figuring out the rules, rather than learning the everyday words first. But then it's not clear what their theory is a theory *of*—it is not, by their concession, a theory of how real children acquire words or how real adults represent them. It may be interesting to indulge in a thought experiment of what an optimal child ought to do if he or she had the entire language to mull over at once. But that exercise would be useful only if the hypothetical child were a good idealization of a real child, and that is far from clear. Why would real children be equipped with an ability to extract intricate chains of rules and arcane word entries if they could never put that ability to use in the real world, and if the net result is the same language as the one they do acquire in the real world? It is more likely that children store words in the mental dictionary in a form that is not radically different in content from what they hear (though it may be more schematic).

Worse, it's not so clear that the thought experiment would come out the way Chomsky and Halle suppose it would. The word pairs that motivate the strange underlying entries, such as *kinesis–kinetic* and *intervene–intervention*, are inkhorn words encountered in writing or in the conversations of literate professionals. Anyone who needs to use these vowel patterns in a new word has the advantage of having seen similar words in print. People who are literate in English have been trained, usually with much weeping and gnashing of teeth, to associate the sounds *ă* and *ā* with the letter *a*, the sounds *ĕ* and *ē* with the letter *e*, and so on, when they learned the alphabet. That means that when speakers have to make a choice from among the short vowels in pronouncing new words such as in *contravene–contravention* or *elide–elision*, they may be guided by their knowledge of the alphabet, not by a naturally acquirable rule of English phonology.

But perhaps the biggest problem is that the Chomsky-Halle-Mohanan theory cannot explain the third kind of similarity running through the irregulars:

the similarities among stems, as in *sting, string, sling, stink, sink, swing,* and *spring*. In their theory the rule of Lowering Ablaut for the participle is connected to these verbs by fiat—the entry for each verb says, "Apply Lowering Ablaut to Me." But then it is an unexplained coincidence that all the verbs are so alike. The list of verbs assigned to the rule could just as easily have been *till–tull, wish–wush, fib–fub,* and *pith–puth*. How can a theory that relentlessly soaks up every droplet of redundancy between stems and pasts, and between the changes applying to one stem and the changes applying to another stem, be so oblivious to the massive redundancy among all the stems undergoing a change? Also, how is the speaker supposed to generalize the rules to new verbs if they are constrained to apply only to the currently stipulated ones?

The obvious way to handle these families is to distill out their common denominator and attach it as a condition to the rule. In the ĭ-ŭ family the ĭ vowel tends to be preceded by a consonant cluster and followed by *ng*. The consonant *ng* can be further analyzed into the features *nasal* (pronounced through the nose) and *velar* (pronounced with the tongue against the soft palate or velum). Perhaps, then, the rule should be "Lower the vowel from ĭ to ŭ if the stem has the pattern *consonant-consonant-ĭ-velar nasal consonant*." Unfortunately, this rule would make errors both of commission and omission. It would falsely include *bring–brought* and *spring–sprang*, which do not change their vowels to ŭ, and it would falsely exclude *stick–stuck* and *spin–spun*, which do. These verbs obviously belong in the class, but each one violates the condition by an eyelash. The *k* of *stick* is not a nasal velar like *ng*, but it is a velar, pronounced at the same place in the mouth. The *n* of *spin* also is not a nasal velar, but it is a nasal, pronounced through the nose, just as *ng* is.[23]

The problem, first pointed out by the linguist Joan Bybee and the psychologist Dan Slobin, is that the irregular clusters are *family resemblance categories*.[24] They don't have strict, all-or-none definitions that specify which verbs are in and which verbs are out. Instead they have fuzzy boundaries and members that are in or out to various degrees depending on how many properties they share with one another. *String* and *sling* are prototypical members of the ĭ-ŭ class, packing into one word all the consonants that are prevalent in the family. *Spin* and *stick* each misses by a different feature; *dig–dug* and *win–won* are farther toward the periphery; and *sneak–snuck, drag–drug, skin–skun,* and *climb–clumb* are in a muzzy zone at the edge where speakers differ as to their acceptability. No rule can cleanly pick out the ĭ-ŭ verbs,

which is why Chomsky, Halle, and Mohanan didn't bother looking for the conditions that triggered each rule but resorted to listing the verbs individually.

The other irregular families work in the same way. For example, *blow–blew, grow–grew,* and *throw–threw* are stereotypical *ow–ew* verbs, but the rule for the class cannot demand that a word conform to the condition *consonant-consonant-ō*. *Know–knew* is in the family but misses the rule by one consonant, *draw–drew* and *fly–flew* miss by a vowel, and *slay–slew* and *crow–crew* are neither clearly in nor clearly out, but muzzy.

Membership in an irregular family is also probabilistic when it comes to people generalizing a pattern to new verbs. Dialectal irregular forms tend to be close in sound to many members of a family. For example, *bring–brang* is close to *sing–sang, ring–rang, spring–sprang, drink–drank,* and *shrink–shrank*, and *write–writ* is close to *bite–bit* and *light–lit*. Fei Xu and I found that children's creative irregulars work the same way. The childhood error *swing–swang* is close to *sing–sang* and all the rest; *sleep–slep* is close to *feed–fed, bleed–bled, meet–met,* and so on.[25] Bybee and Moder even quantified the effect by presenting their adult volunteers with nonsense words that varied in similarity to the typical members of the *ing–ung* family. *Spling* and *skring* fall smack in the middle of the family, and about 80 percent of the participants came up with forms like *splang, splung, skrang,* and *skrung*. *Krink, trig,* and *pling* are less similar, and only about 50 percent of the people suggested *krunk, trug,* or *plang*. *Vin, sid,* and *kib* share only a vowel with the verbs in the family, and only about 20 percent of the people provided forms like *vun, sud,* or *kub*.[26]

Chomsky, Halle, and Mohanan have tweaked rules for maximum performance, but at a steep price. They were forced to make incredible claims about the mental entries of words, and their theory cannot handle the fuzzy and statistical—but psychologically active—patterns of similarity among the verbs undergoing a rule. The irregular patterns are just not very rulelike, and call out for something very different.

~

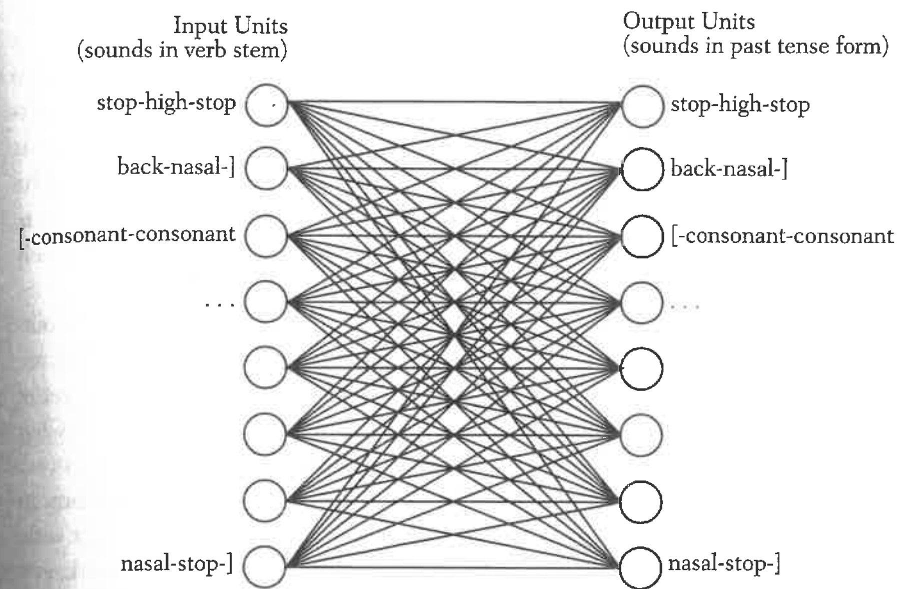
When the psychologists David Rumelhart and James McClelland announced their artificial neural network model of the past tense in 1986, the reaction was sensational.[27] Here was a model with none of the paraphernalia of linguistics—no words, no rules, no modules—but it acquired several hundred regular and irregular past-tense forms, generalized their patterns to new verbs, and made errors such as *breaked* and *comed*, just like children. COMPUTERS MIMIC BRAIN IN TEST, said a headline in the *Chicago Tribune*. A TURNING POINT IN LINGUISTICS, ran the title of a review in the *Times Literary Supplement*.[28] The implications were "awesome," said the reviewer, because "to continue teaching [linguistics] in the orthodox style would be like keeping alchemy alive." Rumelhart and McClelland's model helped to launch a new school of cognitive science known as connectionism or parallel distributed processing, which explains mental processes in terms of networks of interconnected simple units that vaguely resemble neurons (brain cells).[29] Many researchers saw connectionism as a paradigm shift or scientific revolution in the study of the mind.[30] Neural networks also became a fad in artificial intelligence and soon were put to use in picking stocks for mutual funds and controlling expensive Japanese appliances like rice cookers and washing machines.

No one doubts that language is computed by networks of neurons in the brain. Rules—even the pristine, logic-like rules of Chomsky and Halle—are intended as high-level descriptions of processes or structures that are implemented in some way in neural circuitry. The difference between connectionism and generative grammar lies in the *kinds* of mental operations that are thought to be implemented in neural networks. In particular, connectionism differs from generative grammar in the way that associationism differs from symbol manipulation. It lacks combinatorial rules organized into modules, and instead tries to accomplish intelligence using Hume's law of contiguity (if A appears with B, associate them) and his law of resemblance (if C looks like A, let it share A's associations). A neural network that works this way is called a *pattern associator memory* or a *perceptron*.

Here is how Rumelhart and McClelland's model of the past tense works. Despite my clash-of-the-Titans buildup, the model actually shares some important design features with the Chomsky-Halle theory. The input to the model is the sound of a verb stem, and the past tense is computed from it. That is different from a model that computes a past-tense form directly from the meaning of the verb and the concept of pastness. So Rumelhart and McClelland are committed to at least one module—a morphology box—sitting between meaning and sound. As with Chomsky and Halle, a single kind of machinery is charged with computing the past-tense forms of all verbs, regular, irregular, and suppletive (*go–went*); the verbs sit on a continuum of regularity from completely predictable to completely arbitrary. Past-tense forms are composed piecemeal out of miniregularities that are shared among verbs, so that *sleep–slept* combines the vowel change in *feed–fed* and the suffixation in

burn–burnt. Rumelhart and McClelland also import the standard Chomskyan assumption that speech sounds are represented in the mind not as phonemes but as bundles of features such as "voiced" and "nasal."

But everything else is different. Here is the heart of the model, the pattern associator memory:



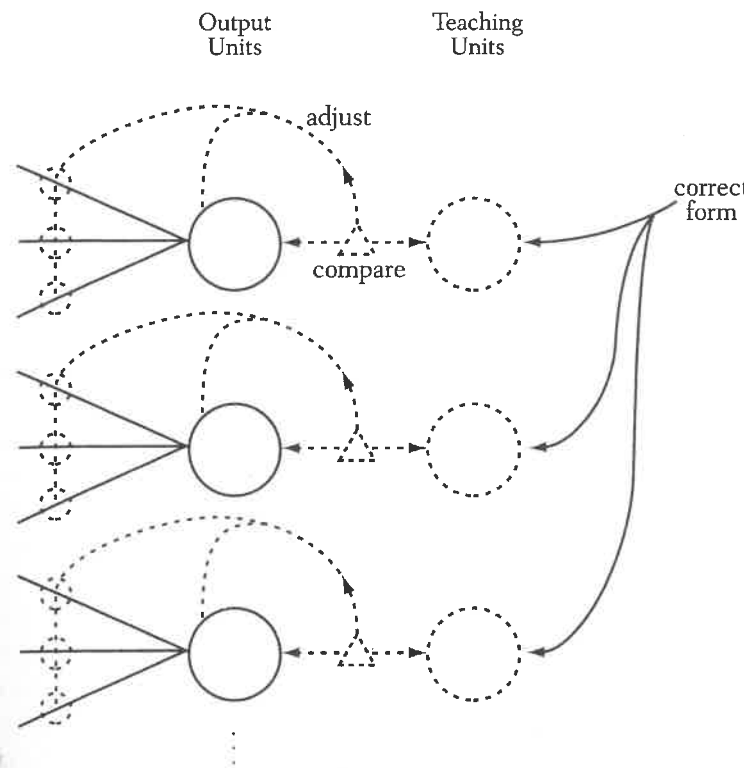
The left-hand column is the input layer, where the verb stem is entered. It contains 460 vaguely neuronlike units, each of which can be either on or off. Each unit represents a tiny stretch of sound that might appear in an English verb, such as a high vowel between two stop consonants, or a back vowel followed by nasal consonant at the end of the word. The beginning and end of a word are symbolized by open and close brackets ("[" and "]"). There are no units for individual verbs; a verb is entered by turning on the units for the sounds it contains. As a result, similar-sounding verbs share representational real estate. Most of the units that are turned on when *shrink* is fed in are also turned on when *drink* is fed in (consonant cluster at the beginning of the word, high vowel between two sonorant consonants, and so on). These units have no idea which word they are currently representing.

The right-hand column has an identical bank of units, and they represent the output of the model: the sound of the past-tense form. Every input is connected to every output by a synapselike connection that can vary in strength, from strongly excitatory (an input signal tends to turn the unit on), to neutral (an input signal has no effect), to strongly inhibitory (an input signal tends to turn the unit off). In effect each connection is a probabilistic microrule that states something like, "If the stem contains a stop consonant followed by a high vowel, the past-tense form is likely to contain a nasal consonant at the end." With 460 input units connected to 460 output units, we have 460 × 460 = 211,600 microrules in all. When an input unit is turned on, it sends a signal down all its lines to the output layer, where the signal is multiplied by the strength of each connection and fed to that output unit. Whether a given output unit turns on depends, in a probabilistic way, on the sum of the signals that feed into it and on its own level of triggerhappiness or *threshold*. The higher the summed signal is above the threshold, the more likely the unit is to turn on; the lower the summed signal is below the threshold, the more likely the unit is to turn off.

In the neonate network the connections have strengths of zero, so the output layer is completely off, regardless of the input. The connections then are changed in a learning procedure, in which the model is "taught" with a set of verbs and their correct past-tense forms. Of course, Rumelhart and McClelland do not actually believe that a schoolmarm has to drill children with verb conjugations. They assume that children, when hearing a past-tense form in their parents' speech, recognize that it *is* the past-tense form of a familiar verb, dredge the verb stem out of memory, feed it into their past tense network, and silently compare their network's output with what they just heard. Skeptics might wonder how a child is supposed to do all this without the benefit of the lexical and grammatical machinery that Rumelhart and McClelland claim to have made obsolete, but let's put that aside for now.

The learning procedure works like this. The correct form from the parents is displayed in a special layer of "teacher" units. The model compares its output, unit by unit, with the correct output (*walked* for *walk*, *came* for *come*, and so on). The model then adjusts the connection strengths a tiny amount up or down depending on the difference (see the figure on the opposite page).

If a unit is off (say, the unit for *ă*), and the teacher says it should be on (because the correct past-tense form is *rang*), the model has to make the input word (*ring*) more likely to turn on that unit in the future. All of the connections from incoming lines that are currently active are strengthened an iota, and the *ă* unit's threshold is lowered an iota, making it more triggerhappy. In contrast, if a unit is on (for example, the unit for ĭ) and the teacher says it should be *off* (because the correct past-tense form is *rang*), the model has to



make the input word *less* likely to turn on that unit in the future. All of the connections from incoming lines that are currently active are weakened an iota (possibly driving the connection down to a negative or inhibitory value), and the unit's threshold is raised an iota, making it less triggerhappy.

The model is trained on a list of verbs and their past-tense forms, presented over and over and over. A given connection will be buffeted up and down by successive verbs in a training run, but eventually it will settle on the strength value that does the best job, in combination with the other connections, of producing correct past-tense forms. The network's knowledge of the various verbs and their past-tense forms is smeared across the 211,600 connection strengths; one cannot point to a circumscribed part of the network that implements a particular word, a particular irregular family, or a regular rule.

Rumelhart and McClelland trained their network on a list of 420 verbs presented 200 times, for a total of 84,000 trials. To everyone's surprise, the model did quite well, computing most of the correct sound stretches for all 420 verbs. That means that a single set of connection strengths was able to convert *look* to *looked*, *seem* to *seemed*, *melt* to *melted*, *hit* to *hit*, *make* to *made*, *sing* to *sang*, even *go* to *went*. Then Rumelhart and McClelland challenged the network with 86 new verbs, which it had not been trained on: a test of generalization or productivity like the *wug*-test, the sine qua non of rules. The model offered the correct past-tense form with *-ed* for about three quarters of the new regular verbs, and made reasonable overgeneralization errors such as *catched* and *digged* for most of the new irregulars.

Even more impressively, the model mimicked some of the tendencies of children as they acquire English. At one point in training it produced errors such as *gived* for verbs that it had previously produced correctly. It also analogized new irregular verbs to families of similar-sounding old irregular verbs; for example, it guessed *cling–clung*, *sip–sept*, *slip–slept*, *bid–bid*, and *kid–kid*. It produced blends such as *gaved* and *stepted* that also occasionally come out of the mouths of children. It was less tempted to tack *-ed* onto an irregular verb from a large family, such as *feel*, than onto an irregular verb from a small family, such as *blow*. And it was bashful about sticking *-ed* onto verbs that already end in *t* or *d*, a common reluctance of human beings that we observed in chapter 2.

Rumelhart and McClelland's pattern associator memory is not made of some miraculous wonder tissue. It works by one trick: Rather than associating a word with a word, it associates the *properties* of a word—its phonological features—with the *properties* of another word, and thereby enjoys automatic generalization by similarity. That is, rather than associating *drink* with *drank*, it associates *dr* with *dr*, *dr* with *rang*, *ring* with *rang*, *ink* with *ank*, and so on. At the same time, it negatively associates *dr* with *nked*, *ink* with *nked*, and so on, inhibiting the incorrect regular form *drinked*.

Crucially, these associations are superimposed across the different words in the training set. When the model trained on *drink* is then trained on *shrink*, it strengthens many of the same connections, such as *ring* with *rang* and *ink* with *ank*. That makes *shrink* easier to learn—most of its connections have been prestrengthened—and it makes subsequent family members, such as *sink*, easier still. It's a short step to generalize to verbs that have not been trained at all, such as *stink*—the *ing–ang* connections have already been strengthened, and the *ing–inged* connections have already been weakened.

The same trick works for the regular verbs: When the model is trained on *walk–walked*, it strengthens connections between *alk* and *alked*, restrengthens them when trained on *talk–talked*, and automatically generalizes them to *stalk–stalked*. The only difference between regular and irregular verbs is that the regulars are more plentiful, more diverse, and more consistent in the patterning of their past-tense forms. With thousands of strong connections conspiring to turn on a *t* or *d* at the end of a word, the model's first tendency will be to output a regular form.

The mainspring of the model, then, is forming associations between features and features, and that duplicates the human habit that embarrassed the words-and-rules theory: generalizing irregular patterns to similar words. The key idea is not original to Rumelhart and McClelland. Associating features with features is inherent to the design of many associationist theories, and goes back to the eighteenth-century English physician and philosopher David Hartley.[31] Hartley pointed out that if the brain represented the properties of an object individually, then Hume's two laws of association—contiguity and similarity—could be pared down to one law, contiguity. Similarity is nothing but shared properties, so associations among properties give you generalization-by-similarity for free. That is, if an association between bread and nourishment is in fact stored in the brain as a set of associations between beigeness, sponginess, savoriness, and nutrition, then when we encounter cake, which is also beige and spongy, "nutritious" pops into mind automatically; no extra device in the brain has to register the fact that bread and cake are similar and make a point of transferring associations from one to the other.

So does traditional linguistics, with its words and rules, have the status of alchemy? Not yet. In 1988 the linguist Alan Prince and I published a paper in the journal *Cognition* that went after the pattern associator model hammer and tongs. We pointed out many facts about human language that the model, and the connectionist approach to language in general, ignored or mishandled.[32] Other trenchant critiques appeared around that time or in the years since.[33] In a recent book the mathematician and former *Scientific American* columnist A. K. Dewdney lumped connectionism with N-rays, cold fusion, and psychoanalysis as case studies of "bad science" in need of debunking.[34] That is unfair, but connectionism *has* been overhyped and its problems as a theory of the mind are real.

Some of the problems might be obvious from the dissection of the language system in chapter 2. First, Rumelhart and McClelland's pattern associator memory is a device that only *produces* past-tense forms. You cannot turn the arrows around and get the model to run backward and *recognize* past-tense forms. Obviously people do both. Not only can we say *walked*, but when we hear *walked* we know it means walk in the past. Children are not separately trained to produce *-ed* and to understand *-ed*. The most straightforward explanation is that they learn rules and lexical entries, a database that can be accessed equally well by a module that sends commands to the tongue and a module that interprets sounds coming in from the ear.

Second, the model computes every detail of the pronunciation of the past-tense form. Yet we saw that many of these details, such as the choice among *-t*, *-d*, and *-ɨd*, are found in fifteen different parts of the language system. Surely they are computed by a single phonology module that is fed by the output of morphology and syntax, not duplicated by an amazing coincidence in fifteen different networks, one for the past tense, one for plurals, and so on.

Third, by forgoing the use of lexical entries and relying entirely on a word's sound to compute its past-tense form, the model cannot tell the difference between two words that have the same sound. It must give them the same past-tense form, and that won't work for soundalike verbs like *ring–rang* and *wring–wrung*, *break–broke* and *brake–braked*, or *meet–met* and *mete–meted*. One might reply that the problem could be fixed by adding a few units to the input that represented the meanings of words, in addition to their sounds. For example, a unit for "striking" could turn on *ang* while a unit for "squeezing" turned on *ung*, differentiating *ring* from *wring*. But as we saw in chapter 2, the meanings of words don't systematically predict their past-tense forms: *Hit*, *strike*, and *slap* are similar in meaning but have different past-tense forms; *take*, *undertake*, and *take a leak* are different in meaning but have the same past-tense forms. It's the raw fact that word 1 is not the same as word 2 which is not the same as word 3 that triggers the different idiosyncratic past-tense forms, and that is the distinction captured by lexical entries. Soundalike words with different plurals and pasts are widespread in English and give rise to many of the quirks that occupy letters to the language mavens—why a baseball player is said to have *flied out* to center field, why the hockey team in Toronto is called *The Maple Leafs*, why the plural of *Walkman* is often *Walkmans*. The answer involves a beautiful design feature of human language that we will explore in chapter 6 and that is quite unlike the knee-jerk associations that drive the Rumelhart-McClelland model.[35]

A fourth problem is that Rumelhart and McClelland had to use some jiggery-pokery to get the model to duplicate children's stages of language development. We will take a closer look at how children really do learn to use and misuse the past tense when we examine language acquisition in chapter 7.[36]

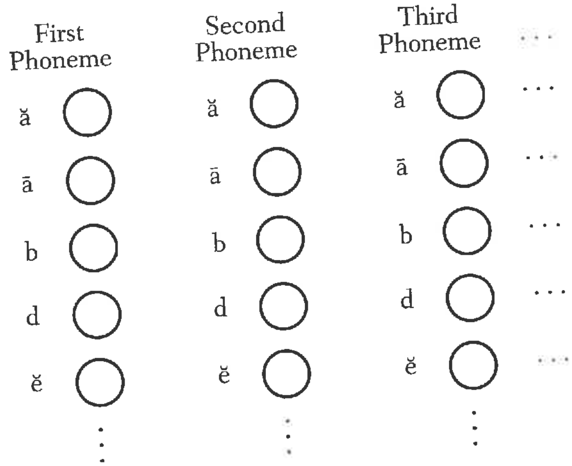
These troubles are all payback for the connectionists' distaste for carving a complex computational problem into a few simpler ones that can be farmed out to mental modules optimized for each. The problems could be solved by building separate networks for morphology, phonology, and the lexicon, much as in traditional linguistics but with the boxes fleshed out as neural networks.[37]

But there is one problem that cannot readily be solved by dividing up the computation into modules. It lies at the very core of the pattern associator model, and diagnoses the main flaw in the centuries-old theory of associationism. The problem could not be more basic: How do you represent an entity made of parts in a fixed arrangement, such as a word? Units can only be on or off; you can't inscribe them with symbols as if they were pads of paper or bytes in a computer. The first solution that comes to mind is to make the units into a phonetic alphabet. Assign one unit to *ā*, one unit to *ă*, one unit to *b*, one unit to *d*, and so on. Then simply turn on the units that spell out the word:

ă ○
ā ○
b ○
d ○
ĕ ○
⋮

But this is a nonstarter. Information about the *order* of phonemes is lost: *pit* would be indistinguishable from *tip*, *Spiro Agnew* from *grow a penis*. If that's all there were to words, you would be solving anagrams every time you opened your mouth.

A better solution is to have an *array* of phoneme units, one bank for the first phoneme in a word, one for the second phoneme, and so on, up to the longest word that a person would ever be called on to remember:



That solves the anagram problem, but it runs up against two new ones. First, how long *is* the longest word that the array should accommodate? Long enough for *antidisestablishmentarianism*, the longest word in standard dictionaries? For *floccinaucinihilipilification*, the longest word in the *Oxford English Dictionary*? What about *great-great-great-grandmother*, *great-great-great-great-great-grandmother*, and so on? There *is* no longest word, so something is wrong with a representation that forces us to decide what it is a priori.

The other problem is that the representation has a bank of units for the first phoneme in a word, a bank of units for the second phoneme, and so on, aligning the words in memory by their first phoneme, that is, left-justifying them. But the human mind does not count off phonemes from left to right when it perceives similarities among words and generalizes accordingly. The most tantalizing generalizations in the irregular past tense system are in the *ing–ang–ung* family, with *ring–rang* and *drink–drank* and *spring–sprang* reinforcing each other and inspiring *fling–flang*, *bring–brang*, and *spling–splang*. But in a left-to-right array, the first three verbs do not overlap at all:

Positions:	1	2	3	4	5	6	7
ring:	*r*	*i*	*ng*				
drink:	*d*	*r*	*i*	*ng*	*k*		
spring:	*s*	*p*	*r*	*i*	*ng*		

Nothing that has been learned for *ring*, such as that it has a past tense with *ang*, will transfer to *drink* or *spring*; the two words are no more similar than *bird* and *clam*. But people *must* find them similar, for the irregular system is rife with families of words that don't align properly at their left edges, such as *dive, drive, strive*, and most obviously, prefixed forms such as *stand, withstand*, and *understand* or *come, become*, and *overcome*. The same problem fouls up any generalization that depends on the *ends* of words, and we know there is a huge one: the choice of *-t*, *-d*, or *-ɨd*, which hinges on whether the last phoneme is voiced, unvoiced, or a *t* or *d*. The last phoneme in a left-to-right representation can be position 2 (for the verb *add*), position 3 (for *ask*), position 4 (for *risk*), and so on, all the way up to position 23 for *floccinaucinihilipilify* and beyond. A left-to-right representation would have to learn how to pronounce the suffix separately for every length of word.

Rumelhart and McClelland must have recognized this problem, because they came up with a creative alternative: The units stand for things they called *Wickelphones*, named after the psychologist Wayne Wickelgren who first conceived them.[38] A Wickelphone is a sequence of three phonemes, like *ipt* or *str*. English has about forty phonemes, and if we add the special symbols "[" and "]" for the beginning and end of a word, there are about 67,000 possible Wickelphones, each needing a unit. By representing a word by its Wickelphones, one sidesteps both the anagram problem and the left-alignment problem. For example, *strip* contains the Wickelphones *ip]*, *rip*, *str*, *tri*, and *[st*. You don't have to worry about their order, because they snap together in only one way: *[st* at the beginning, then *str*, then *tri*, and so on. And the Wickelphones for *rip* overlap the Wickelphones for *strip*, so their representations are similar, just as the human mind perceives them. (Rumelhart and McClelland in fact wanted to represent words in terms of their Chomsky-Halle-esque features, rather than their phonemes, so a unit actually stood for three *features* in a row, such as stop-high-stop or voiced-unvoiced-voiced—a Wickelfeature.)

It may seem hard to believe that the simple act of registering a word in an unstructured bank of units is a near-insoluble problem, but it is: Wickelphones, though ingenious, don't work either. The human mind cares about single phonemes and the features that compose them, and the Wickelphone has submerged them into unbreakable chains of three-in-a-row. For example, *silt* and *slit* have no Wickelphones in common. The first dissolves into *[si, sil, ilt*, and *lt]*, the second into *[sl, sli, lit*, and *it]*. But people clearly hear them as similar, as we see in historical changes in English such as *brid* becoming *bird* and *thrid* becoming *third*. Worse, in some languages the Wickelphone cannot represent certain words at all. The Australian aboriginal language Oykangand has word *algal* meaning "straight" and a word *algalgal* meaning "ramrod straight," and they are made up of identical Wickelphones: *alg, al], gal, lga,* and *[al*:

Words:	[algal]	[algalgal]
Wickelphones:	[al	[al
	alg	alg
	lga	lga
	gal	gal
	al]	*alg* (already used)
		lga (already used)
		gal (already used)
		al]

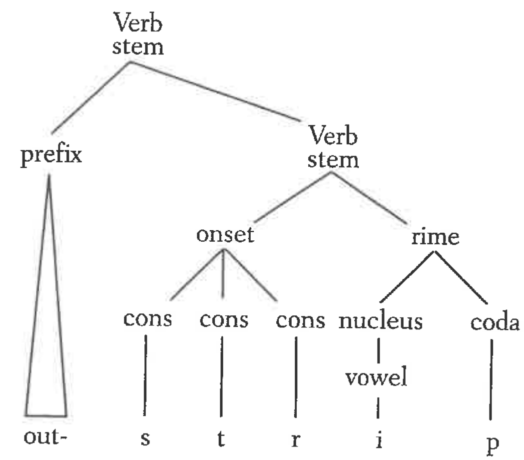
Since units are either on or off, they have no way of representing *two* of something, and the Wickelphone theory therefore incorrectly predicts that Oykangand speakers should not exist (nor the speakers of many other languages; words like these are not uncommon).[39]

Also, a theory of how the mind represents things should predict what the mind finds easy and what it finds hard. The easy tasks should be computed by simple operations on the representation, the hard tasks by lengthy sequences of operations. Here, too, the Wickelphone makes the wrong prediction. When linguists explain to their classes how human languages use some kinds of rules and not others, they almost always use the same example: that no language has a rule that flips a word to its mirror-image, say, forming the plural by converting *tip* to *pit*, *gum* to *mug*, and *dog* to *god*. But a Wickelphone-to-Wickelphone network can do exactly that, and quite easily: strengthen every connection between a Wickelphone *ABC* in the input and its mirror-image Wickelphone *CBA* in the output, and weaken all the other connections.

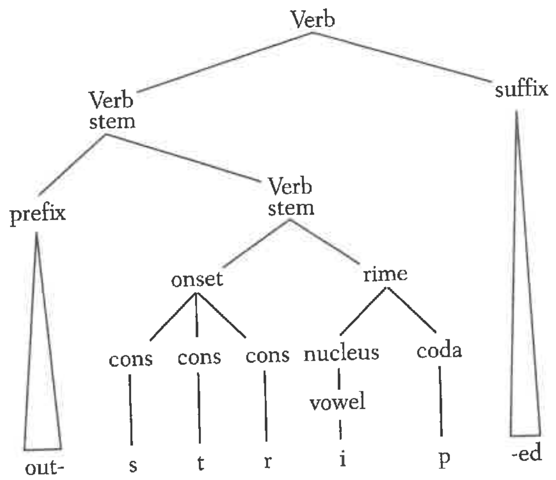
Not only is mirror-reversal easy, but it is no harder to learn than the simplest conceivable relation between input and output: copying the stem verbatim, which involves strengthening the connections between *ABC* and *ABC*. The only difference between mirror-image reversal and verbatim copying is that we, the theorists peering into the model, can read the unit labels and see that *ABC* goes to *CBA* in one case and to *ABC* in the other. But the model cannot read its own node labels; all it cares about is the consistency of the input-output relations, and they are the same in both cases. Likewise, all kinds of crazy rules, such as replacing all *as* with *bs*, all *bs* with *cs*, all *cs* with *ds*, and so on, are as easy to learn as copying the input to the output.

This is not just a quibble; it explains an embarrassing lapse in the performance of Rumelhart and McClelland's model. The model was mute when asked for the past tenses of simple but somewhat unusual-sounding words, like *jump, pump, warm,* and *trail*. And it garbled several others, turning *squat* into *squakt*, *tour* into *toureder*, and *mail* into *membled*. The lapses are puzzling to *us* because intuitively nothing could be simpler than copying a stem over to the past-tense form before adding *-ed*. But a pattern associator memory has no placeholder called "stem" that *can* be copied, and no operation to do the copying. All it does is associate sounds with sounds, and if the training set happens to be missing words with certain combinations of sounds such as *-ump* or *-ail*, the model will be at a loss, and will sit in silence or cough up a hairball of bits and pieces that are vaguely associated with the sounds it *has* been trained on.[40]

All the problems go away if you bring back the rationalist theory that the mind manipulates symbols organized into hierarchical structures by rules. A verb such as *to outstrip* might be represented something like this:



The phonemes are held in their correct order by a treelike scaffolding that embodies the morphological structure of the word (how it is built out of stems, prefixes, and suffixes) and the phonological structure of its parts (how they are built out of chunks like onsets, rimes, vowel nuclei, consonants and vowels, and ultimately, features). The similarity to other words such as *strip, restrip, trip, rip,* and *tip* fall mechanically out of the fact that they have identical subtrees, such as an identical "stem" or an identical "rime." And computing the regular past-tense form is nothing but attaching a suffix next to the symbol "verb stem":



It doesn't matter whether the underbrush dangling beneath the "verb stem" symbol is *walk, outstrip, jump, pump*, or *bftsplk*—if you have a mental symbol "verb stem" and know how to put a suffix next to it, the entire vocabulary of verb stems lies waiting at your feet. Finally, since a tree structure is built out of recursive rules (for example, "a stem can combine with a prefix to form a new stem"), no length limit needs to be set beforehand, and words of any length such as *re-outstrip* or *great-great-great-grandmother* can be represented.

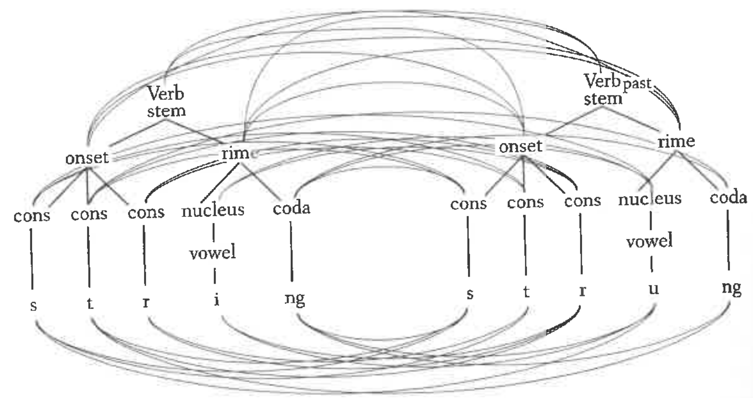
Symbolic trees require fancier neural hardware than the smooth purée of units that are popular among connectionists, but those models hardly do justice to the brain anyway. Recently, a few neural network modelers have shown how hierarchical trees can be implemented in more organized neural networks.[41] One conjecture is that the periodic rhythms of neural firing, long downplayed in neuroscience, serve as the glue that binds together the units that represent an abstract *slot* in a tree and the units that represent its *content*. For example, when the units for the "coda" slot fire at twenty times a second and the units for *p* fire in synchrony with them, also at twenty times a second, the system as a whole knows that the coda is *p*. Simultaneously, the units for "nucleus" can be firing thirty times a second, and the units for *i* can be firing in synchrony thirty times a second, and the system knows that the nucleus is *i*. The units for "coda," "nucleus," *p*, and *i* are all active at the same time, but the system doesn't get confused and think that *i* is in the coda, because the shared firing patterns link each sound to its slot. That theory may or may not be right,

but I mention it to show that abstract symbols and complex structure are not incompatible with plausible neural network models.

After Alan Prince and I took apart the pattern associator model, the linguists breathed a sigh of relief because they thought they didn't have to learn neural network modeling after all, and the connectionists dropped it like a hot potato. So it is ironic that Prince and I are probably the model's biggest fans. It does, after all, explain a major phenomenon that rule theories ignore, and it accounts for not one but several aspects of children's language development. In comparison, the twenty-five connectionist models of the past tense that have been devised in reply to our critique have been disappointments, not one of them anywhere near as ambitious as the original.[42] Many sidestep the Wickelphone problem by using a Dick-and-Jane version of English that contains only monosyllabic words made of a consonant, a vowel, and a consonant, such as *walk* and *run*. Others implicitly concede that words are composed of symbols for stems and symbols for affixes and don't even bother computing a past-tense form. They merely select from an innate menu of five or six units that stand for the five or six suffixes or vowel changes in the language. Some *other* mechanism then has to apply the suffix or vowel change to the stem to get an actual past-tense form. That unmentioned mechanism, of course, is what we call a *rule*. Many modelers beef up the network with an intervening layer of units hidden between the input and the output layers, but direct benchmark tests find little or no improvement.[43] Each of the inventors has added a different patch that narrowly fixes some problem that Prince and I pointed out—what programmers call a hack or a kluge—but none defends his brainchild as an actual theory of how that part of the mind works. And no one has made an empirical prediction or accounted for several kinds of data in the way that Rumelhart and McClelland did.

~

One phenomenon, two models, both explaining too much to be completely wrong, both too flawed to be completely right. Prince and I have proposed a hybrid in which Chomsky and Halle are basically right about regular inflection and Rumelhart and McClelland are basically right about irregular inflection. Our proposal is simply the traditional words-and-rules theory with a twist. Regular verbs are computed by a rule that combines a symbol for a verb stem with a symbol for the suffix. Irregular verbs are pairs of words retrieved from the mental dictionary, a part of memory. Here is the twist: Memory is not a list of unrelated slots, like RAM in a computer, but is associative, a bit like the Rumelhart-McClelland pattern associator memory. Not only are words linked to words, but bits of words are linked to bits of words. The bits are not Wickelphones, of course, but substructures like stems, onsets, rimes, vowels, consonants, and features, perhaps something like this:



Furthermore, the nodes of one word (such as *string*) overlap the same nodes in other words (such as *sling, stick, stink*, and *swim*). As a result, irregular verbs show the kinds of associative effects found in a connectionist pattern associator. People find families of similar irregular verbs easier to store and recall because these verbs repeatedly strengthen their shared associations. And people occasionally generalize the irregular patterns to new, similar verbs, because the new verbs contain material that already had been associated with the pattern from the old verbs.

Prince and I were not the first to modify the words-and-rules theory in this way. Many generative linguists have been uncomfortable with the Chomsky-Halle ethos of using industrial-strength rules to account for everything that is systematic in language. Mark Aronoff, Joan Bresnan, Ray Jackendoff, Rochelle Lieber, Andrew Spencer, and others have suggested that language uses two kinds of rules: true rules that speakers generalize freely, and *lexical redundancy rules* that merely capture patterns of similarity among words stored in memory.[44] A memory system in which patterns of similarity are registered and occasionally generalized is simply a pattern associator memory, and Rumelhart and McClelland have given us a sketch of how one might work.

The modified words-and-rules theory may sound like a sappy attempt to get everyone to make nice and play together, but it makes a strong prediction. The prediction is that regular and irregular inflection are psychologically, and ultimately, neurologically distinguishable. But how could they be distinguished if both involve patterns that people can generalize? The answer is that irregular inflection depends on *memorized* words or forms *similar* to them, but regular inflection can apply to *any* word, regardless of whether the word is readily retrievable from memory. Regular inflection has that power because it is computed by a mental operation that does not *need* access to the contents of memory, namely, a symbol-processing operation or rule, which applies to any instance of the symbol "verb." The evidence will be woven through the rest of the book as we explore how words are used in conversation and in reading, how new words are created, how children learn their mother tongue, how language is organized in the brain, and whether the languages of the world conform to a universal design. We will see how in dozens of cases of language use that have nothing in common except a failure of access to memory, irregular patterns are disabled but the regular rule works fine.

If the modified words-and-rules theory is correct, it would have a pleasing implication for the centuries-old debate between associationism and rationalism: Both theories are right, but they are right about different parts of the mind.