Learning Sound Patterns

To get a sense of what's really involved in learning language, cast your mind back to what it was like before you knew any words at all of your native tongue. Well, wait … since you obviously can't do that, the best you can do is to recall any experiences you may have had learning a second language at an age old enough to remember what the experience was like (or a third or fourth language, if you were lucky enough to learn more than one language as a tot). If these memories involve learning a language in a classroom setting, they turn out to be a useful point of departure for our purposes, especially to highlight the striking difference between how you learned language in the classroom, and how you learned it as a newborn initiated into your native language.

In a foreign language classroom, it's usual for the process to kick off with a teacher (or textbook) translating a list of vocabulary items from the new language into your native language. You then use a small but growing vocabulary to build up your knowledge of the language. You begin to insert words into prefabricated sentence frames, for example, and eventually you build sentences from scratch. This is simply not an approach that was available to you as an infant because then you had no words in any language that could be used as the basis of translation. Worse, you didn't even know what words were, or where words began or ended in the stream of speech you were listening to. You were basically swimming in a sea of sound, and there wasn't a whole lot anyone could do in the way of teaching that would have guided you through it.

If you were to have the unusual experience of learning a second language by simply showing up in a foreign country and plunging yourself into the language as best you could, without the benefit of language courses or tourist phrase books, that would be a bit closer to what you faced as an infant.
But still, you have advantages as an adult that you didn’t have as an infant. You are much more sophisticated in your knowledge of the world, so you’re not faced with learning how to describe the world using language while you’re trying to figure out what the world is like. And your intellect allows you to be much more strategic in how you go about getting language samples from the speakers of that language; you can, for example, figure out ways to ask speakers about subtle distinctions—like whether there are different words for the concepts of cat and kitten, or how to interpret the difference in similar expressions, such as screw up and screw off. Not to mention that your motor skills allow you to pantomime or point to objects as a way to request native speakers to produce the correct words for you.

Deprived of many of the possible learning strategies that older people use, it might make sense that babies would postpone language learning until they develop in other areas that would help support this difficult task. And, given the fact that most babies don’t start producing recognizable words until they’re about a year old, and that they take quite a bit longer than that to string sentences together, it might seem that that’s exactly what does happen. But in fact, babies begin learning their native language from the day they’re born, or even earlier; it turns out that French babies tested within 4 days of birth could tell the difference between French and Russian, and sucked more enthusiastically on a pacifier when hearing French (Mehler et al., 1988). On the other hand, infants born into other linguistic households, such as Arabic, German, or Chinese, did not seem to be able to tell the difference between French and Russian speech, nor did French household babies seem to notice the difference between English and Italian. This study indicates that babies in utero can begin to learn something about their native language. Obviously, this can’t be the result of recognizing actual words and their meanings, since in utero babies have no experience of the meanings that language communicates. Rather, it suggests that even through the walls of the womb and immersed in amniotic fluid, babies learn something about the patterns of sounds in the language they hear.

Humans are unlike honeybees and certain species of songbirds, which are genetically programmed for a specific type of bee dance or birdsong. The speech and accent of a child born of French parents but raised from infancy in the United States will usually be indistinguishable from that of a child born to U.S. parents and raised in that country. This linguistic flexibility reflects the fact that humans are powerful learning machines.

In this chapter, we’ll look at what young children need to learn about the sounds of their language—and the sound system of any language is an intricate, delicately patterned thing. Not only does it have its own unique collection of sounds, but it has different “rules” for how these sounds can be combined into words. For example, even though English contains all the individual sounds of a word like ptak, it would never allow them to be strung in this order, though Czech speakers do so without batting an eye. And no word in Czech can ever end in a sound like “g,” even though that consonant appears in abundance at the beginnings and in the middles of Czech words. English speakers, on the other hand, have no inhibitions about uttering a word like dog.

In fact, the sound pattern of a language is a complex code that infants manage to crack and mostly master within the first couple of years of life. A magnificent amount of learning happens within the first few months of birth. Long before they begin to produce words (or even really show that they understand their meanings), babies can:

1. Differentiate their native language from other languages.
2. Have a sense of how streams of sound are carved up into words.
3. Give special attention to distinctions in sounds that will be especially useful for signaling different meanings (for example, the distinction between “p” and “b” sounds; switch the “p” in pat to a “b” sound, and you get a word with a different meaning).

4. Figure out how sounds can be “legally” combined into words in their language.

Babies also develop many other nifty skills. Perhaps the only people who deserve as much admiration as these tiny, pre-verbal human beings are the scientists who study the whole process. Unlike foreign language teachers who can test students’ mastery of a language via multiple choice exams and writing samples, language researchers have to rely on truly ingenious methods for probing the infant mind. In this chapter, you’ll get a sense of the accomplishments of both groups: the very young who crack the sound code, and the scientists who study their feats.

4.1 Where Are the Words?

A stunningly ineffective way to learn language would be to simply memorize the meaning of every complete sentence you’ve ever heard, never bothering to break the sentence down into its component parts. In fact, if you took this route (even if you could actually memorize thousands upon thousands of complete sentences in the form of long strings of sound), you wouldn’t really be learning language at all. No matter how many sentences you’d accumulated in your memory stash, you’d constantly find yourself in situations where you were required—but unable—to produce and understand sentences that you’d never encountered before. And without analyzing language in terms of its component, reusable parts, learning a sentence like *My mother’s sister has diabetes* would be no help at all in understanding the very similar sentence *My father’s dog has diabetes*. You’d treat the relationship in sound between these as purely coincidental, much as you would the relationship between the similar-sounding words *battle* and *cattle*; each one would have to be memorized completely independently of the other. As you saw in Chapter 2, the aspect of language that lets us combine meaningful units (like words) to produce larger meaningful units (like phrases or sentences) is one of the universal properties of human language, and one that gives it enormous expressive power.

Fundamentally, learning language involves figuring out which sounds clump together to form basic units, and learning how these units in turn can be combined with other units—which is why foreign language instruction for beginners puts so much emphasis on learning lists of words. One of the infant’s earliest tasks, then, is to figure out which strings of sounds form these basic units—no trivial accomplishment. In talking to their babies, parents are not nearly as accommodating as Spanish or German textbooks, and they rarely speak to their children in single-word utterances (about 10 percent of the time). This means that babies are confronted with speech in which multiple words are sewn seamlessly together, and they have to figure out all on their own where the edges of words are. And unlike written language, where words are clearly separated by spaces (at least in most writing systems), spoken language doesn’t present convenient breaks in sound to isolate words. To get an intuitive feel for what this task might feel like to a baby, see how you fare in Web Activity 4.1.

WEB ACTIVITY 4.1

Finding word boundaries  In this activity, you’ll hear speech in several different languages, and you’ll be asked to guess where the word boundaries might be.
The head-turn preference paradigm is an invaluable tool that’s been used in hundreds of studies of infant cognition. It can be used with babies as young as about 4.5 months, and up to about 18 months. After this age, toddlers often become too fidgety to reliably sit still through this experimental task. Applied to speech perception, the technique is based on two simple principles:

1. Babies turn their heads to orient to sounds.
2. Babies spend more time orienting to sounds that they find interesting.

In a typical experiment, the baby sits on the lap of a parent or caregiver who is listening to music over a set of headphones; this prevents the adult from hearing the experimental stimuli and either purposely or inadvertently giving cues to the baby. A video camera is set up to face the child, recording the baby’s responses, and an unseen observer monitors the experiment by watching the baby on video. (The observer can’t hear the stimulus sounds and is usually not aware of which experimental condition the child has been assigned to, though the observer does control the sequence of events that occur.) A flashing light mounted next to the video camera and straight ahead in the child’s view can be activated to draw the child’s attention to a neutral point before any of the experimental stimuli are played. The stimuli of interest are then played on two speakers, mounted on the left and right walls (see Figure 4.1).

Each experiment usually consists of a familiarization phase and a test phase. There are three goals for the familiarization phase. The first is simply to have the infant become familiar with the sound stimuli. In some cases, if the purpose of the study is to find out whether babies will learn something about new sounds, the sounds played during the familiarization phase might consist of the new stimuli to be learned. The second goal is to train the baby to expect that sounds can come from the speaker on either the left or the right wall. The third goal of the familiarization phase is to tightly lock together the head-turn behavior to the infant’s auditory attention. Babies tend to look in the direction of a sound that holds their attention anyway, but this connection can be strengthened by flashing a light.

Figure 4.1 A testing booth set up for the head-turn preference paradigm. The baby sits on the caregiver’s lap, facing the central panel. The observer looks through a small window or one-way mirror to note the baby’s head turns. (Adapted from Nelson et al., 1995.)
Learning Sound Patterns

Probing infants’ knowledge of words

Babies begin to produce their first words at about a year or so, but they start to identify word breaks at a much younger age than that. In fact, the whole process is under way by 6 or 7 months. Scientists can’t exactly plop a transcript of speech in front of a baby, hand them a pencil, and ask them to mark down where the word breaks are. So how is it possible to peer into infants’ minds and determine whether they are breaking sentences down into their component parts?

In studying the cognitive processes of infants, researchers have to content themselves with a fairly narrow range of infant behaviors as a way to measure hidden psychological mechanisms. So, when they decide to study infants of a particular age, they need to have a clear sense of what babies can do at that point in their development—and more specifically, which behaviors reflect meaningful cognitive activity. It turns out that a great deal of what we now know about infant cognition rests on one simple observation: when babies are faced with new sounds or images, they devote their attention to them differently than when they hear or see old, familiar sounds or images. And at the age of 6 or 7 months, one easy way to tell if a baby is paying attention to something is if she swivels her head in its direction to stare at it; the longer she keeps her gaze oriented in its direction, the longer she’s paying attention to that stimulus. New sounds and sights tend to draw attention differently than familiar ones, and babies will usually orient to novel versus familiar stimuli for different lengths of time—sometimes they’re more interested in something that’s familiar to them (a sort of “Hey, I know what that is!” response), but sometimes they prefer the novelty of the new stimulus.

These simple observations about the habits of babies gave birth to a technique that psycholinguists now commonly use, called the head-turn preference paradigm (see Method 4.1). This technique compares how long babies keep their heads turned toward different stimuli, taking this as a measure of their attention. (If the target stimulus is a sound, it’s usually coupled with a visual stimulus such as a light or a dancing puppet in order to best elicit the

in the location of the speaker before each sound, and by making sure that sounds during the familiarization phase are played only for as long as the baby looks in the direction of the sound. This signals to the child that if she wants to continue hearing a sound, she needs to be looking in its direction. After all these goals have been achieved, the baby is ready for the test phase.

During the test phase, the sounds of interest are played on either the left or the right speaker, and the baby’s head-turn behavior is recorded by the video camera for later coding. Researchers then measure how long the baby spends looking in the direction of each sound, and these responses are averaged over stimulus type.

Sometimes researchers are interested in which sounds the infants prefer to listen to—do they prefer a female voice to a male voice, for example, or do they prefer to listen to sounds of their own language over an unknown language? Other times, the researchers are only interested in whether the infants discriminate between two categories of stimuli, and it doesn’t matter which category is preferred. For instance, in a learning experiment, any distinction in looking times for familiar versus new stimuli should be an indication of learning, regardless of whether babies prefer to listen longer to the new or the familiar sounds. And in fact, it turns out that there isn’t a clear preference for either new or familiar sounds—at times, babies show more interest in sounds that they recognize, and at other times, they show more interest in completely novel ones. The preferences seem to depend somewhat on the age of the child and just how often they’ve heard the familiar sounds (overly familiar sounds might cause a baby to become bored with them).

METHOD 4.1 (continued)

head-turn preference paradigm An experimental framework in which infants’ speech preference or learning is measured by the length of time they turn their heads in the direction of a sound.
head-turn response.) But what makes the method really powerful is that it can leverage the measure of looking time as a way to test whether or not the babies taking part in the study have learned a particular stimulus. For instance, let’s say we give babies a new word to listen to during a familiarization phase. At some later time, during a test phase, we can see whether their looking times when hearing this word are different from those for a word they’ve never heard before. If babies spend either more or less time looking at the previously presented word than they do at a completely new word, this suggests that they’ve learned something about the first word and now treat it as “familiar.” On the other hand, if they devote equal amounts of looking time to both words, it suggests that they haven’t learned enough about the previously heard word to differentiate it from a completely novel word.

The head-turn preference paradigm has been used—for example, by Peter Jusczyk and Richard Aslin (1995)—to tackle the question of whether babies have learned where word breaks occur. Here’s how: During the familiarization phase of that study, the baby participants heard a series of sentences that contained a target word, say, bike, in various different positions in the sentence:

- His bike had big black wheels.
- The girl rode her big bike.
- Her bike could go very fast.
- The bell on the bike was really loud.
- The boy had a new red bike.
- Your bike always stays in the garage.

During the test phase, the researchers measured how long the infants were interested in listening to repetitions of the target word bike, compared with a word (say, dog) that they hadn’t heard during the familiarization phase. To first get the baby’s attention before the test word was played, a flashing light appeared above the loudspeaker the word was to come from. Once the baby looked in this direction, the test word repetitions began to play. When the baby’s interest flagged, causing him to look away from the loudspeaker, this was noted by a researcher, and the baby was scored for the amount of time spent looking in the direction of the loudspeaker. Jusczyk and Aslin found that overall, 7.5-month-old babies spent more time turning to the speaker when it played a familiar word (bike) than when it played an unfamiliar word (dog). This might not seem like a tremendous feat to you, but keep in mind that the babies must somehow have separated the unit bike from the other sounds in the sentences during the familiarization phase in order to be able to match that string of sounds with the repeated word during the test phase. Six-month-old babies didn’t seem to have this ability yet.

This study shows that by the tender age of 7.5 months, babies seem to be equipped with some ability to separate or segment words from the speech stream—but it doesn’t tell us how they manage to come by these skills, or what information they rely on to decide where the words are. Since Jusczyk and Aslin’s initial study, dozens of published articles have explored the question of how babies pull this off. We’ll investigate several ways that they might begin to crack the problem.

**Familiar words break apart the speech stream**

Here’s one possibility. Remember that babies hear single-word utterances only about 10% of the time. That’s not a lot, but it might be enough to use as a starting point for eventually breaking full sentences apart into individual words. It
may be that babies can use those few words they do hear in isolation as a way to build up a small collection of known word units. These familiar words can then serve as anchoring points for breaking up the speech stream into more manageable chunks. For example, imagine hearing this procession of sounds in a fictional language:

bankiritubendudifin

Any guesses about what the word units are? The chances of getting it right are not very high. But suppose there are two words in this stream that you’ve heard repeatedly because they happen to be your name (“Kiri”) and the name for your father (“Dudi”). You may have learned them because these are among the few words that are likely to be uttered as single words quite often, so they’re especially easy to recognize. Perceptually, they’ll leap out at you. If you’ve learned a foreign language, the experience of hearing sentences containing a few familiar words may be similar to the very early stages of learning a new language as a baby; you might have been able to easily pull out just one or two familiar words from an otherwise incomprehensible sentence. With this in mind, imagine hearing:

ban- kiri- tuben- dudi- fin

Now when you hear the names kiri and dudi, their familiarity allows you to pull them out of the speech stream—but it might also provide a way to identify other strings of sound as word units. It seems pretty likely that ban and fin are word units too, because they appear at the beginning and end of the utterance and are the only syllables that are left over after you’ve identified kiri and dudi as word units. So now, you can pull out four stand-alone units from the speech stream: kiri, dudi, ban, fin. You don’t know what these last two mean, but once they’re firmly enough fixed in your memory, they might in turn serve as clues for identifying other new words. So tamfinatbankirisan can now be pulled apart into:

tam- fin- at- ban- kiri- san

The residue from this new segmentation yields the probable units tam, at, and san, which can be applied in other sentences in which these units are combined with entirely new ones.

In principle, by starting with a very few highly familiar words and generating “hypotheses” about which adjacent clumps of sound correspond to units, an infant might begin to break down streams of continuous sound into smaller pieces. In fact, experimental evidence shows that babies are able to segment words that appear next to familiar names when they’re as young as 6 months old, suggesting that this strategy might be especially useful in the very earliest stages of speech segmentation. (Remember that the Jusczyk and Aslin study showed that 6-month-olds did not yet show evidence of generally solid segmentation skills.) This was demonstrated in an interesting study led by Heather Bortfeld (2005). Using the head-turn preference paradigm, the researchers showed that even 6-month-olds could learn to segment words such as bike or feet out of sentences—but only if they appeared right next to their own names (in this example, the baby subject’s name is Maggie) or the very familiar word Mommy:

Maggie’s bike had big, black wheels.
The girl laughed at Mommy’s feet.

That is, when babies heard these sentences during the familiarization phase, they later spent more time looking at loudspeakers that emitted the target word bike or feet than at speakers that played an entirely new word (cup or dog)—this
shows they were treating bike and feet as familiar units. But when the sentences in the familiarization phase had the target words bike and feet right next to names that were not familiar to the child (for example, The girl laughed at Tommy’s feet), the babies showed no greater interest in the target words than they did in the new words cup or dog. In other words, there’s no evidence that the infants had managed to pull these words out of the stream of speech when they sat next to unknown words. At this age, then, it seems that babies can’t yet segment words out of just any stream of speech, but that they can segment words that appear next to words that are already very familiar.

**Discovering what words sound like**

Relying on familiar words to bust their way into a stream of sound is just one of the tricks that babies have in their word-segmentation bag. Another trick has to do with developing some intuitions about which sounds or sequences of sounds are allowed at the beginnings and ends of words, and using these intuitions as a way to guess where word boundaries are likely to be.

To get a more concrete feel for how this might work, try pronouncing the following nonsense sentence in as “English-like” a way as you can, and take a stab at marking the word boundaries with a pencil (hint: there are six words in the sentence):

Banriptangbowpkesternladfloop.

If you compare your answer with those of your classmates, you might see some discrepancies, but you’ll also find there are many similarities. For example, chances are, no one proposed a segmentation like this:

Ba-nri-ptangbow-pkester-nladfl-oop.

This has to do with what we think of as a “possible” word in English, in terms of the sequence of sounds it’s allowed to contain. Just because your language includes a particular sound in its inventory doesn’t mean that sound is allowed to pop up just anywhere.

Languages have patterns that correspond to what are considered “good” words as opposed to words that look like the linguistic equivalent of a patched-together Frankenstein creature. For example, suppose you’re a marketing expert charged with creating a brand new word for a line of clothing, and you’ve decided to write a computer program to randomly generate words to kick-start the whole process. Looking at what the computer spat out, you could easily sort out the following list of words into those that are possible English-sounding words, and those that are not:

- ptangb
- sastashak
- roffo
- lululeming
- spimton
- ndela
- skrs
- srbridl

What counts as a well-behaved English word has little to do with what’s actually pronounceable—you might think that it’s impossible to pronounce sequences of consonants like pt, nd, and dl, but actually you do it all the time in words like riptide, bandage, bed linen—it’s just that in English, these consonants have to straddle word boundaries or even just syllable boundaries. (Remember, there are no actual pauses between these consonants just by virtue of their belonging to different syllables or words). You reject the alien words like ptangb or ndela in your computer-generated list, not because it takes acrobatic feats of
your mouth to pronounce them, but because you have ingrained word templates in your mind that you’ve implicitly learned, and these words don’t match those mental templates. These templates differ from one language to another and are known as phonotactic constraints.

Using English phonotactic constraints to segment another language, though, could easily get you into trouble, especially if you try to import them into a language that’s more lenient in allowing exotic consonant clusters inside its words. For instance, skrs and ndela are perfectly well-formed pronunciations of words in Czech and Swahili, respectively. So, trying to segment speech by relying on your English word templates would give you non-optimal results. You can see this if you try to use English templates to segment the following three-word stream of Swahili words:

nipemkatenzuri

You might be tempted to do either of the following:

nipem-katen-zuri
nip-emkat-enzuri

but the correct segmentation is:

nipe-mkate-nzuri

And while languages like Czech and Swahili are quite permissive when it comes to creating consonant clusters that would be banned by the rules of English, other languages have even tighter restrictions on clusters than English does. For instance, “sp” is not a legal word-initial cluster in Spanish (which is why speakers of that language often say “espanish”). (You can see some more examples of how languages apply different phonotactic constraints in Box 4.1.)

It turns out that by 9 months of age, babies have some knowledge of the templates for proper words in their language. Using the head-turn preference paradigm, researchers led by Peter Jusczyk (1993) have shown that American babies orient longer toward strings of sounds that are legal words in English (for example, cubeb, dudgeon) than they do to sequences that are legal Dutch words but illegal words in English (zampjes, vlatke). Dutch 9-month-olds show exactly the opposite pattern. This suggests that they’re aware of what a “good” word of their language sounds like. And, just as neither you nor your classmates suggested that speech should be segmented in a way that allows bizarre words like ptangb or nladf, at 9 months of age, babies can use their phonotactic templates to segment units out of speech (Mattys & Jusczyk, 2001).

You might have noticed in Web Activity 4.2 that there was another bit of information that might have helped you segment words, in addition to clues about phonotactic constraints. In that exercise, English-like stress patterns were also present. In English, stress tends to alternate, so within a word, you usually get an unstressed syllable sitting next to a stressed one: reTURN, BLACKmail, invigorate. (In some other languages, such as French, syllables are more or less evenly stressed.) English words can follow either a trochaic stress pattern, in which the first syllable is stressed (as in BLACKmail), or an iambic stress pattern, in which the first syllable is unstressed (as in reTURN). But as it turns out, it’s not an equal-opportunity distribution, and trochaic words far outnumber iambic words (on the order of 9 to 1 by some estimates). Chances are, you subconsciously made use of this knowledge in your segmentation answers in Web Activity 4.2. If babies have caught on to this pattern in
As languages go, English is reasonably loose in allowing a wide range of phonotactic templates. English allows consonants to gather together in sizable packs at the edges of words and syllables. For example, the single-syllable word *splints* has the structure CCCVCCC (where C = consonant and V = vowel). Many other languages are far more restrictive. To illustrate, the Hebrew, Hawaiian, and Indonesian languages allow only the following syllable structures:

<table>
<thead>
<tr>
<th>Language</th>
<th>CV</th>
<th>CVC</th>
<th>CVCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hebrew</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Hawaiian</td>
<td>V</td>
<td>VC</td>
<td>CV</td>
</tr>
<tr>
<td>Indonesian</td>
<td>V</td>
<td>CV</td>
<td>CVC</td>
</tr>
</tbody>
</table>

In addition to broadly specifying the consonant-vowel structure of syllables, languages have more stringent rules about which consonants can occur where. For example, in English, /rp/ is allowed at the end of the word, but not at the beginning; the reverse is true for the cluster /pr/. Some constraints tend to recur across many languages, but others are highly arbitrary. The table gives a few examples of possible and impossible clusters that can occur at the beginnings of words and syllables in several languages.

<table>
<thead>
<tr>
<th>Language</th>
<th>/kn/</th>
<th>/skw/</th>
<th>/sb/</th>
<th>/vr/</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>German</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>French</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Italian</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

To get a feel for how speech segmentation might be affected by language-specific phonotactic constraints, try listing all the possible ways to break down the following stream of sounds into word units that are legal, depending on whether your language is English, German, French, or Italian:

bakniskweriavrosbamanuesbivriknat

To get at generalizations like these, babies must already have segmented some word units, held them in memory, and “noticed” (unconsciously, of course) that they follow certain patterns. All this from speech that rarely has any words that stand alone without the added confusion of adjacent speech sounds. Earlier we speculated that maybe isolated familiar words like the baby’s name serve as the very first units; these first words then act as a wedge for segmenting out other words, allowing the baby to build up an early collection of words they might also be able to use to make guesses about how words are segmented from speech. And indeed, they do: by 7.5 months, babies have no trouble slicing words with a trochaic pattern (like *DOCtor*) from the speech stream—but when they hear an iambic word like *guiTAR* embedded in running speech, they don’t recognize it. In fact, if they’ve heard *guitar* followed by the word *is*, they behave as if they’ve segmented *TARis* as a word (Jusczyk et al., 1999).

But if you’ve been paying very close attention, you might have noticed a paradox: in order for babies to be able to use templates of permissible words to segment speech, they already have to have some notion of a word—or at the very least, they have to have some notion of a language unit that’s made up of a stable collection of sounds that go together and can be separated from other collections of sounds. Otherwise, how can they possibly have learned that “ft” can occur as a sequence of sounds in the middle or at the end of a word unit, but not at its beginning? It’s the same thing with stress patterns: How can babies rely on generalizations about the most likely stress patterns for words in their language unless they’ve already analyzed a bunch of words?

To get at generalizations like these, babies must already have segmented some word units, held them in memory, and “noticed” (unconsciously, of course) that they follow certain patterns. All this from speech that rarely has any words that stand alone without the added confusion of adjacent speech sounds.
of word units. Once this set gets large enough, the baby can learn some useful generalizations that can then accelerate the whole process of extracting additional new words from running speech. In principle, this is plausible, given that babies can use familiar words like names to figure out that neighboring bunches of sounds also form word units. But this puts quite a big burden on those very first few words. Presumably, the baby has managed to identify them because they’ve been pronounced as single-word utterances. But as it happens, parents are quite variable in how many words they produce in isolation—some produce quite a few, but others are more verbose and rarely provide their children with utterances of single words. If this were the crucial starting point for breaking into streams of speech, we might expect babies to show a lot more variability in their ability to segment speech than researchers typically find, with some lagging much further behind others in their segmentation abilities.

Luckily, youngsters aren’t limited to using familiar, isolated words as a departure point for segmentation—they have other, more flexible and powerful tricks up their sleeves. Researchers have discovered that babies can segment streams of sounds from a completely unfamiliar language after as little as two minutes of exposure, without hearing a single word on its own, and without the benefit of any information about phonotactic constraints or stress patterns. How they manage this accomplishment is the topic of the next section.

### 4.2 Infant Statisticians

*Tracking transitional probabilities: The information is out there*

In a seminal study, Jenny Saffran and colleagues (1996) familiarized 8-month-old infants with unbroken 2-minute strings of flatly intoned, computer-generated speech. The stream of speech contained snippets such as:

```
bidakupadotigolabubidaku
```

Notice that the sounds are sequenced so that they follow a repeating consonant-vowel structure. Because English allows any of the consonants in this string to appear either at the beginnings or the ends of syllables and words, nothing about the phonotactic constraints of English offers any clues at all about how the words are segmented, other than that the consonants need to be grouped with at least one vowel (since English has no words that consist of single consonants). For example, the word *bidakupa* could easily have any of the following segmentations, plus a few more:

- bi-dak-upa
- bid-aku-pa
- bid-ak-u-pa
- bi-da-kup-a
- bid-ak-up-a
- bidaku-pa
- bi-dakup-a
- bid-akup-a
- bidak-upa

If this seems like a lot, keep in mind that these are the segmentation possibilities of a speech snippet that involves just four syllables; imagine the challenges involved in segmenting a two-minute-long continuous stream of speech. This is precisely the task that Saffran and colleagues inflicted on the babies they studied.

In the Saffran et al. study, though, the stream of sound that the babies listened to during their familiarization phase was more than just a concatenation of consonant-vowel sequences. The stimuli were created in a way that repre-
Experimenters create an artificial “language” of four “words”: bidaku, golabu, padoti, dutaba.

**Familiarization phase**
Infant hears each “word” repeated 45 times in random order, in an unbroken 2-minute synthesized speech stream:

```
```

**Test phase**
Loudspeakers present infant either a “real” word: bidaku, golabu

or a sequence of syllables with parts of two words: dakupa, buduta

**Results**

<table>
<thead>
<tr>
<th></th>
<th>Mean looking times for 8-month-olds</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Real” words</td>
<td>6.77 seconds</td>
</tr>
<tr>
<td>Part-words</td>
<td>7.60 seconds</td>
</tr>
</tbody>
</table>

Figure 4.2 In this study, Saffran and colleagues prepared stimuli that amount to a miniature artificial language of four “words,” each word consisting of three consonant-vowel syllables. Infants then heard an uninterrupted, 2-minute stream of random combinations of the four words. The researchers noted how much attention the babies paid to the four “words” from the familiarization phase and compared it with the attention the babies paid to three-syllable sequences that also occurred in the speech but that straddled “word” boundaries (part-words). (Adapted from Saffran et al., 1996.)

**artificial language** A “language” that is constructed to have certain specific properties for the purpose of testing an experimental hypothesis: strings of sounds correspond to “words,” which may or may not have meaning, and whose combination may or may not be constrained by syntactic rules.

WEB ACTIVITY 4.3

Segmenting “words”: You be the baby! In this activity, you’ll hear 2 minutes of stimuli from an artificial language very similar to that used by Saffran et al. (1996). You’ll be asked to distinguish “words” from “non-words” to see if you, too, can manage to segment speech. (Ideally, you should attempt this exercise before you read any further.)
leagues suggested that babies were acting like miniature statisticians analyzing the speech stream, and were keeping track of **transitional probabilities (TPs)** between syllables—this refers to the likelihood that one particular syllable will be followed by another specific syllable. Here’s how such information would help to define likely word units: Think of any two syllables, say, a syllable like *ti* and a syllable like *bay*. Let's say you hear *ti* in a stream of normal English speech. What are the chances that the very next syllable you hear will be *bay*? It’s not all that likely; you might hear it in a sequence like *drafty basement* or *pretty baby*, but *ti* could just as easily occur in sequences that are followed by different syllables, as in *T-bone steak*, *teasing Amy*, *teenage wasteland*, *Fawlty Towers*, and many, many others.

But notice that *ti* and a *bay* that follows it don’t make up an English word. It turns out that when a word boundary sits between two syllables, the likelihood of predicting the second syllable on the basis of the first is vanishingly small. But the situation for predicting the second syllable based on the first looks very different when the two syllables occur together **within** a word. For example, take the sequence of syllables *pre* and *ti*, as in *pretty*. If you hear *pre*, as pronounced in this word, what are the chances that you’ll hear *ti*? They’re much higher now—in this case, you’d never hear the syllable *pre* at the end of a word, so that leaves only a handful of words that contain it, dramatically constraining the number of options for a following syllable. Generally, the transitional probabilities of syllable sequences are much higher for pairs of syllables that fall within a word than for syllables that belong to different words. This is simply because of the obvious fact that words are **units** in which sounds and syllables clump together to form a fairly indivisible whole. Since there’s a finite number of words in the language that tend to get used over and over again, it stands to reason that the TPs of syllable sequences within a word will be much higher than the TPs of syllable pairs coming from different words.

How does all this help babies to segment speech? Well, if the little tykes can somehow manage to figure out that the likelihood of hearing *ti* after *pre* is quite high, whereas the likelihood of hearing *bay* after *ti* is low, they might be able to respond to this difference in transitional probabilities by “chunking” *pre* and *ti* together into a word-like unit, but avoid clumping *ti* and *bay* together.

Here’s the math: The transitional probability can be quantified as $P(Y|X)$, that is, the probability that a syllable $Y$ will occur given that the syllable $X$ has just occurred. This is done by looking at a sample of a language and dividing the frequency of the syllable sequence $XY$ by the frequency of the syllable $X$ combined with any syllable:

$$TP = P(Y|X) = \frac{\text{frequency (XY)}}{\text{frequency (X)}}$$

In the study by Saffran and her colleagues, the only cues to word boundaries were the transitional probabilities between syllable pairs: within words, the transitional probability of syllable pairs (e.g., *bida*) was always 1.0, while the transitional probability for syllable pairs across word boundaries (e.g., *kupa*) was always 0.33.

That babies can extract such information might seem like a preposterous claim. It seems to be attributing a whole lot of sophistication to tiny babies. You might have even more trouble swallowing this claim if you, a reasonably intelligent adult, had trouble figuring out that transitional probabilities were the relevant source of information needed to segment the speech in Web Activity 4.3 (and you wouldn’t have been alone in failing to come up with an explanation for how the speech might be segmented). How could infants possibly manage to home in on precisely these useful statistical patterns when you failed to see...
them, even after studying the speech sample and possibly thinking quite hard about it?

Though it might seem counterintuitive, there’s a growing stack of evidence that a great deal of language learning “in the wild”—as opposed to in the classroom—actually does involve extracting patterns like these, and that babies and adults alike are very good at pulling statistical regularities out of speech samples, even though they may be lousy at actually manipulating math equations. As we’ll see later, sensitivity to statistical information applies not just to segmenting words from an unfamiliar language, but also to learning how sounds make patterns in a language, how words can be combined with other words, or how to resolve ambiguities in the speech stream. The reason it doesn’t feel intuitive is that all of this knowledge is implicit and can be hard to access at a conscious level. For example, you may have done reasonably well in identifying the “words” in the listening portion of Web Activity 4.3, even if you had trouble consciously identifying what information you were using in the analysis task. (Similarly, you may have had easy and quick intuitions about the phonotactic constraints of English but worked hard to articulate them.) That is, you may have had trouble identifying what it is you knew and how you learned it, even though you did seem to know it. It turns out that the vast majority of our knowledge of language has this character—throughout this book, you’ll be seeing many more examples of a seeming disconnect between your explicit, conscious knowledge of language and your implicit, unconsciously learned linguistic prowess.

Being able to track transitional probabilities gives infants a powerful device for starting to make sense of a running river of speech sounds. It frees them up from the need to hear individual words in isolation in order to learn them, and it solves the problem of how they might build up enough of a stock of words to serve as the basis for more powerful generalizations about words—for example, that in English, words are more likely to have a trochaic stress pattern than an iambic one, or that the consonant cluster “ft” can’t occur at the beginning of a word, though it can occur at its end. Ultimately, these generalizations, once in place, may turn out to be more robust and useful for word segmentation than transitional probabilities are. At times, such generalizations might even conflict with the information provided by transitional probabilities. Eventually, infants will need to learn how to attend to multiple levels of information, and to weight each one appropriately.

Is statistical learning a specialized human skill for language?

We’ve now seen some of the learning mechanisms that babies can use to pull word-like units out from the flow of speech they hear, including keeping track of various kinds of statistical regularities. Now let’s step back and spend a bit of time thinking about how these learning mechanisms might connect with some of the bigger questions laid out in Chapters 2 and 3.

Much of Chapter 2 focused on the questions of whether language is unique to humans, and on whether certain skills have evolved purely because they make efficient language use possible. In that chapter, I emphasized that it was impossible to think of language as a monolithic thing; learning and using language involve an eclectic collection of skills and processes. Since we’ve now begun to isolate what some of those skills might look like, we can ask a much more precise question: Do non-humans have the ability to segment speech by keeping track of statistical regularities among sounds?

If you found yourself surprised and impressed at the capacity of babies to statistically analyze a stream of speech, you may find it all the more intriguing to learn that as a species, we’re not alone in this ability. In a 2001 study,
researchers led by Marc Hauser replicated the earlier studies by Saffran and colleagues; their subjects were not human infants, however, but cotton-top tamarins, a species of tiny monkey (Figure 4.3A). The monkeys heard about 20 minutes’ worth of the same artificial language used by Saffran and her colleagues in their human experiments—four three-syllable “words” like tupiro and bidaku, strung together in random order and with no pauses between syllables or word units. Afterward, the monkeys went through a test phase, just as the human babies had, in which they heard a sequence of three syllables over a loudspeaker. As in the study with humans, these syllables could correspond to either a word from the artificial language (e.g., bidaku) or a part-word in which two syllables of an actual word were combined with a third syllable (for example, tibida). The researchers looked at how often the monkeys turned to face the speaker when they heard the test stimulus. They found that the monkeys distinguished between the two types of syllable sequences and oriented toward the speaker more often when they heard a part-word than when they’d heard a word (Figure 4.3B).

These results address the question of whether only humans are equipped to learn to chop up speech streams by paying attention to statistical regularities: apparently not. But the experiment doesn’t fully get at the question of whether this statistical ability is exclusively enmeshed with our capacity for language. Cotton-top tamarins do have a fairly complex system of sequential calls, and maybe both their system of vocalizations and their statistical abilities reflect precursors of a fully linguistic system like ours. To better test for the connection or disconnection between statistical learning and language, it might make sense to study a species that shows no signs of having moved toward a human-like system of communication. Juan Toro and Josep Trobalón (2005) did just this in their study of speech segmentation in rats, using the same artificial language that previous researchers had used with human infants and cotton-top tamarins. They found that rats, too, were able to use statistical regularities in the speech to learn to differentiate between “words” and “non-words” of that language.

This suggests that picking up on statistical cues may be a very general cognitive skill—one that’s not monopolized by species that have language or language-like communication systems, and one that might be useful in domains other than language. If that’s so, then we should find that humans don’t just pull this trick out of their hats for the purpose of learning language, but that we can also make use of it when confronted with very different kinds of stimuli. And indeed, this turns out to be true. Humans are capable of picking up on
The head-turn preference paradigm (see Method 4.1) is a clever behavioral method that has allowed researchers to test infants' knowledge without requiring any sophisticated responses or behaviors. Nevertheless, it does require babies to have developed the neck muscles that are needed to turn their heads in response to a stimulus. It also requires the babies to sustain full consciousness for reasonable periods of time. This makes it challenging to study the learning skills of newborn babies, with their floppy necks and tendency to sleep much of the time when they’re not actively feeding. But as you saw in Chapter 3, ERPs (event-related potentials) can be used to probe the cognitive processes of people in a vegetative state, bypassing the need for any meaningful behavior at all in response to stimuli. Could the same method be used to assess the secret cognitive life of newborns?

Tuomas Teinonen and colleagues (2009) used ERP methods to test whether newborns can pick up on the transitional probabilities of syllables in a sample of speech. Their subjects were less than 2 days old, and they listened to at least 15 minutes of running speech consisting of ten different three-syllable made-up words randomly strung together. After this 15-minute “learning” period, the researchers analyzed the electrical activity in the babies’ brains. Because the ERPs of newborn babies are less wildly variable if measured during sleep, the researchers limited the analysis to brain activity that was monitored during active sleep—which turned out to represent 40%–80% of the hour-long experiment.

ERP activity was compared for each of the three syllables of the novel “words.” The logic behind this comparison was that, since the first syllable for any given “word” was less predictable (having a lower transitional probability) than the second and third syllables, it should show heightened brain activity compared with the other two syllables. Figure 4.4 shows the results of the study, which indicate a region of enhanced negative activity for the first of the three syllables. Similar results have since shown that newborns can also track the transitional probabilities of tones (Kudo et al., 2011). These remarkable studies reveal that the ability to pull statistical regularities from the auditory world is a robust skill that’s available to humans from the very first moments after birth.

Figure 4.4 Results from Teinonen et al., Experiment 1. ERP activity at two recording sites (F3 and C3) shows enhanced negativity. In each pair of panels, the syllables are aligned so that each syllable’s onset corresponds to 0. The shaded areas show the region where there is a statistically significant difference between the first syllable (S1) and the second and third syllables (S2 and S3). (Adapted from Teinonen et al., 2009.)
regularities within stimuli as diverse as musical tones (Saffran et al., 1999) and visual shapes (Fiser & Aslin, 2001). It appears, then, that one of the earliest language-related tasks that a baby undertakes rests on a pretty sturdy and highly general cognitive skill that we share with animals as we all try to make sense of the world around us. Indeed, it’s likely that as humans, we literally have this ability from birth (see Box 4.2).

But that’s not the end of the story. Just because individuals of different species can track statistical regularities across a number of different domains doesn’t necessarily mean that the same kinds of regularities are being tracked in all these cases. In fact, Toro and Trobalón found that rats were able to use simple statistical cues to segment speech but weren’t sensitive to some of the more complex cues that have been found to be used by human infants. And there may also be some subtle distinctions in the kinds of cues that are used in dealing with language, for example, as opposed to other, non-linguistic stimuli. These more nuanced questions are taken up in Digging Deeper at the end of this chapter.

4.3 What Are the Sounds?

How many distinct sounds are there in a language?

You might think that having to figure out where the words are in your language is hard enough. But in fact, if we back up even more, it becomes apparent that babies are born without even knowing what sounds make up their language. These sounds, too, have to be learned. This is not as trivial as it seems. As an adult whose knowledge of your language is deeply entrenched, you have the illusion that there’s a fairly small number of sounds that English speakers produce (say, about 40), and that it’s just a matter of learning what these 40 or so sounds are. But in truth, English speakers produce many more than 40 sounds.

Here’s an example: Ever since your earliest days in school, when you were likely given exercises to identify sounds and their corresponding letter symbols, you learned that the words tall and tree begin with the same sound, and that the second and third consonants of the word potato are identical. But that’s not exactly right. Pay close attention to what’s happening with your tongue as you say these sounds the way you normally would in conversational speech, and you’ll see that not all consonants that are represented by the letter t are identical. For example, you likely said tree using a sound that’s a lot like the first sound in church, and unless you were fastidiously enunciating the word potato, the two “t” sounds were not the same. It turns out that sounds are affected by the phonetic company they keep. And these subtle distinctions matter. If you were to cut out the “t” in tall and swap it for the “t” in tree, you would be able to tell the difference. The resulting word would sound a bit weird. The sound represented by the symbol t also varies depending on whether it’s placed at the very beginning of a syllable, as in tan, or is the second member of a consonant cluster, as in Stan.

Not convinced? Here’s some playing with fire you’re encouraged to try at home. Place a lit match a couple of inches away from your mouth and say the word Stan. Now say tan. If the match is at the right distance from your mouth (and you might need to play around with this a bit), it will be puffed out when you say tan, but not when you say Stan. When you use “t” at the beginning of a syllable, you release an extra flurry of air. You can feel this if you hold your palm up close to your mouth while saying these words.

These kinds of variations are in no way limited to the “t” sound in English; any and all of the 40-odd sounds of English can be and are produced in a variety of different ways, depending on which sounds they’re keeping company
with. Suddenly, the inventory of approximately 40 sounds has mushroomed into many more.

Not only do the surrounding sounds make a difference to how any given sound is pronounced, but so do things like how fast the speaker is talking; whether the speaker is male or female, old or young; whether the speaker is shouting, whispering, or talking at a moderate volume; and whether he or she is talking to a baby or a friend at a bar, or is reading the news on a national television network.

And yet, despite all this variation, we do have the sense that all “t” sounds, regardless of how they’re made, should be classified as representing one kind of sound. This sense goes beyond just knowing that all of these “t” instances are captured by the orthographic symbol t or T. More to the point, while swapping out one kind of “t” sound for another might sound weird, it doesn’t change what word has been spoken. Not like swapping out the “t” in ten for a “d” sound, for example. Now, all of a sudden, you have a completely different word, den, with a completely different meaning. This means that not all sound distinctions are created equal. Some change the fundamental identity of a speech sound, while others are the speech equivalent of sounds putting on different outfits depending on which other sounds they’re hanging out with, or what event they happen to be at.

When a sound distinction has the potential to actually cause a change in meaning, that distinction yields separate phonemes. But when sound differences don’t fundamentally change the identity of a speech unit, we say they create different allophones of the same phoneme. You know that sound distinctions create different phonemes when it’s possible to create minimal pairs of words in which changing a single sound results in a change in meaning. For example, the difference between “t” and “d” is a phonemic distinction, not an allophonic distinction, because we get minimal pairs such as ten, den; toe, doe; and bat, bad (see Table 4.1).

Our impression is that the difference between the sounds “t” and “d” is a big one, while the difference between the two “t” sounds in tan and Stan is very slight and hard to hear. But this sense is merely a product of the way we mentally categorize these sounds. Objectively, the difference between the sounds in both pairs is close to exactly the same, and as you’ll see later on, there’s evidence that we’re not deaf to the acoustic differences between allophones—but we’ve mentally amplified the differences between phonemes, and minimized the differences between allophones.

**A catalogue of sound distinctions**

To begin to describe differences among speech sounds in a more objective way, it’s useful to break them down into their characteristics. This turns out to be

<table>
<thead>
<tr>
<th>TABLE 4.1 Examples of minimal word pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>pad/bad</td>
</tr>
<tr>
<td>tap/tab</td>
</tr>
<tr>
<td>fan/van</td>
</tr>
</tbody>
</table>

*In English, the presence of minimal word pairs that differ only with respect to a single sound shows that those sounds (boldface type) are distinct phonemes. Be sure to focus on how the words sound rather than on how they are spelled.
Learning Sound Patterns

PLACE OF ARTICULATION Consonants are typically made by pushing air out of the lungs, through the larynx and the vocal folds (often called the “vocal cords,” although the term “folds” is much more accurate) and through the mouth or nose (Figure 4.5). The vocal folds, located near the top of the larynx, are a pair of loosely attached flaps that vibrate as air passes through them; these vibrations produce sound waves that are shaped into different speech sounds by the rest of the vocal tract. (To hear your vocal folds in action, try first whispering the syllable “aahh,” and then utter it as you normally would—the “noise” that’s added to the fully sounded vowel comes from the vocal fold vibration, or phonation.) To create a consonant sound, the airflow passing through the vocal tract has to be blocked—either partially or completely—at some point above the larynx. The location where this blockage occurs has a big impact on what the consonant sounds like. For example, both the “p” and “t” sounds are made by closing the air off at the lips, while “t” is made by closing it off at the little ridge just behind your teeth, or the alveolar ridge. A sound like “k,” on the other hand, is made by closing the air off at the back of the mouth, touching the palate with the back of the tongue rather than its tip. And these are really the only significant differences between these sounds. (As you’ll see in a moment, along the other two sound dimensions, “p,” “t,” and “k” are all alike.) Moving from lips to palate, the sounds are described as bilabial for “p,” alveolar for “t,” and velar for “k.” Other intermediate places exist as well, as described next and summarized in Figure 4.6.

Figure 4.5 The human vocal tract, showing the various articulators. Air from the lungs passes through the larynx and over the vocal folds, making the folds vibrate and thus producing sound waves. The tongue, lips, and teeth help form this sound into speech. The place of articulation refers to the point at which the airflow becomes obstructed; for example, if airflow is briefly cut off by placing the tongue against the alveolar ridge, a sound would be said to be alveolar; a sound made by obstructing airflow at the velum would be velar.

quite easy to do, because speech sounds vary systematically along a fairly small number of dimensions. For example, we only need three dimensions to capture most consonants of English and other languages: place of articulation, manner of articulation, and voicing.

<table>
<thead>
<tr>
<th>Place of articulation</th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Interdental</th>
<th>Alveolar</th>
<th>Alveopalatal</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>k</td>
<td>g</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>f</td>
<td>v</td>
<td>θ</td>
<td>s</td>
<td>j</td>
<td></td>
<td></td>
<td>h</td>
</tr>
<tr>
<td>Affricative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral liquid</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retroflex liquid</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glide</td>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.6 A chart of the consonant phonemes of Standard American English. In this presentation, the sounds are organized by place of articulation, manner of articulation, and voicing. (From the International Phonetic Association.)

vocal folds Also known as “vocal cords,” these are paired “flaps” in the larynx that vibrate as air passes over them. The vibrations are shaped into speech sounds by the other structures (tongue, alveolar ridge, velum, etc.) of the vocal tract.

phonation Production of sound by the vibrating vocal folds.

bilabial Describes a sound that is produced by obstructing airflow at the lips.
MANNER OF ARTICULATION  As mentioned, the airflow in the vocal tract can be obstructed either completely or partially. When the airflow is stopped completely somewhere in the mouth, you wind up producing what is known as a stop consonant. Stop consonants come in two varieties. If the air is fully blocked in the mouth and not allowed to leak out through the nose, you have an oral stop—our old friends “p,” “t,” and “k.” But if you lower the velum (the soft tissue at the back of the roof of your mouth; see Figure 4.5) in a way that lets the air pass through your nose, you’ll produce a nasal stop, which includes sounds like “m,” “n,” and the “ŋ” sound in words like sing or fang. You might have noticed that when your nose is plugged due to a cold, your nasal stops end up sounding like oral stops—“my nose” turns into “by dose” because no air can get out through your stuffed-up nose.

But your tongue is capable of more subtlety than simply blocking airflow entirely when some part of it is touched against the oral cavity. It can also narrow the airflow in a way that produces a turbulent sound—such as “s” or “f” or “z.” These turbulent sounds are called fricatives. If you squish an oral stop and a fricative together, like the first and last consonants in church or judge, you wind up with an affricate.

Or, you can let air escape over both sides of your tongue, producing what are described as liquid sounds like “l” or “r,” which differ from each other only in whether the blade (the front third) of your tongue is firmly planted against the roof of your mouth or is bunched back.

Finally, if you obstruct the airflow only mildly, allowing most of it to pass through the mouth, you will produce a glide. Pucker your lips, and you’ll have a “w” sound, whereas if you place the back of your tongue up toward the velum as if about to utter a “k” but stop well before the tongue makes contact, you’ll produce a “y” sound.

VOICING The last sound dimension has to do with whether (and when) the vocal folds are vibrating as you utter a consonant. People commonly refer to this part of the human anatomy as the “vocal cords” because, much like a musical instrument (such as a violin or cello) that has strings or cords, pitch in the human voice is determined by how quickly this vocal apparatus vibrates. But unlike a cello, voice isn’t caused by passing something over a set of strings to make them vibrate. Rather, sound generation in the larynx (the “voice box”) involves the “flaps” of the vocal folds (see Figure 4.5), which can constrict either loosely or very tightly. Sound is made when air coming up from the lungs passes through these flaps; depending on how constricted the vocal folds are, you get varying amounts of vibration, and hence higher or lower pitch. Think of voice as less like a cello and more like air flowing through the neck of a balloon held either tightly or loosely (though, in terms of beauty, I’ll grant that the human voice is more like a cello than like a rapidly deflating balloon).

Vowels, unless whispered (or in certain special situations), are almost always produced while the vocal folds are vibrating. But consonants can vary. Some, like “z,” “v,” and “d,” are made with vibrating vocal folds, while others, like “s,” “f,” and “t,” are not—try putting your hand up against your throat just above your Adam’s apple, and you’ll be able to feel the difference.

Oral stops are especially interesting when it comes to voicing. Remember that for these sounds, the airflow is completely stopped somewhere in the mouth when two articulators come together—whether two lips, or a part of the tongue and the roof of the mouth. Voicing refers to when the vocal folds begin to vibrate relative to this closure and release. When vibration happens just about simultaneously with the release of the articulators (say, within about 20
It’s not likely that a YouTube video of someone reciting a random list of words from the *Oxford English Dictionary* would spread virally—the act is just not that interesting. But many people are rightly riveted by the skills of virtuoso beatbox artists. Beatboxing is the art of mimicking musical and percussive sounds, and during their performances beatboxers routinely emit sounds with names like 808 snare drum roll, brushed cymbal, reverse classic kick drum, bongo drum, and electro scratch. When you see them in action, what comes out of their mouths seems more machine-like than human.

And yet, when you look at how these sounds are actually made, it becomes clear that the repertoire of beatbox sounds is the end result of creatively using and recombining articulatory gestures that make up the backbone of regular, everyday speech. In fact, the connection between speech and beatboxing is so close that, in order to notate beatbox everyday sounds, artists have used the International Phonetic Alphabet as a base for Standard Beatbox Notation.

Want to know how to make the classic kick drum sound? On the website Humanbeatbox.com, beatboxer Gavin Tyte explains how. First, he points out:

In phonetics, the classic kick drum is described as a bilabial plosive (i.e., stop). This means it is made by completely closing both lips and then releasing them accompanied by a burst of air.

To punch up the sound, Tyte explains, you add a bit of lip oscillation, as if you were blowing a very short “raspberry.” Step by step, in Tyte’s words:

1. Make the “b” sound as if you are saying “b” from the word bogus.
2. This time, with your lips closed, let the pressure build up.
3. You need to control the release of your lips just enough to let them vibrate for a short amount of time.

The classic kick drum sound (represented as “b” in Standard Beatbox Notation) can be made as a voiced or voiceless version. Embellishments can be added: you can add on fricative sounds (“bsh,” “bs,” or “bf”), or combine the basic sound with a nasal sound (“bng,” “bm,” or “bn”).

What sounds really impressive, though, is when a beatbox artist combines actual words with beatbox rhythms—it sounds as if the artist is simultaneously making speech sounds and non-speech sounds. But this is really a trick of the ear. It’s not that the artist is making two sounds at the same time, but that he’s creating a very convincing auditory illusion in which a single beatbox sound swings both ways, being heard both as a musical beatbox sound and as a speech sound. The illusion relies on what’s known as the phonemic restoration effect. Scientists have created this effect in the lab by splicing a speech sound like “s” out of a word such as *legislature*, completely replacing the “s” sounds with the sound of a cough. Listeners hear the cough, but they also hear the “s” as if it had never been removed. This happens because, based on all the remaining speech sounds that really are there, the mind easily recognizes the word *legislature* and fills in the missing blanks (more on this in Chapter 7). In order for the illusion to work, though, the non-speech sound has to be acoustically similar to the speech sound. So, part of a beatboxer’s skill lies in knowing which beatbox sounds can double as which speech sounds. Though many beatboxers have never taken a course in linguistics or psycholinguistics, they have an impressive body of phonetic knowledge at their command.

From a performance standpoint, skilled beatboxers display dazzling articulatory gymnastics. They keep their tongues leaping around their mouths in rapid-fire rhythms, and coordinate several parts of their vocal tracts all at the same time. But newbies to the art shouldn’t be discouraged. It’s certainly true that learning to beatbox takes many hours of practice. But when you think about it, the articulatory accomplishment is not all that different from what you learned to do as an infant mastering the sounds of your native language, and learning to put them all together into words. As you saw in Box 2.4, most infants spend quite a bit of time perfecting their articulatory technique, typically passing through a babbling stage beginning at about 5 months of age, in which they spend many hours learning to make human speech sounds. In the end, learning to beatbox may take no more practice than the many hours you were willing to put in learning how to talk—just think back to the hours you spent in your crib, taking your articulatory system out for a spin, and babbling endlessly at the ceiling.

[**phonemic restoration effect**](#) Auditory illusion in which people “hear” a sound that is missing from a word and has been replaced by a non-speech sound. People report hearing both the non-speech sound and the “restored” speech sound at the same time.
milliseconds) as it does for “b” in the word ban, we say the oral stop is a voiced one. When the vibration happens only at somewhat of a lag (say, more than 20 milliseconds), we say that the sound is unvoiced or voiceless. This labeling is just a way of assigning discrete categories to what amounts to a continuous dimension of voice onset time (VOT), because in principle, there can be any degree of voicing lag time after the release of the articulators.

You might have noticed that all of the consonants listed in Figure 4.6 end up being different phonemes. That is, it’s possible to take any two of them and use them to create minimal pairs, showing that the differences between these sounds lead to differences in meaning, as we saw in Table 4.1. But that table shows you only a phonemic inventory of English sounds, not the full range of how these sounds are produced, in all their glorious allophonic variety, when each phoneme trots out its full wardrobe.

Phonemes versus allophones: How languages carve up phonetic space

Now that you have a sense of the dimensions along which sounds vary, I owe you a convincing account of why the differences between phonemes are often no bigger than the differences between allophones. (I’m talking here about their differences in terms of pure sound characteristics, not your mental representations of the sounds.)

Let’s first talk about the differences in the “t” sounds in tan and Stan. Remember that extra little burst of air when you said tan? That actually comes from a difference in voice onset time. That is, there’s an extra-long lag between when you release your tongue from your alveolar ridge and when your vocal folds begin to vibrate, perhaps as long as 80 milliseconds (ms). You get that extra puff because more air pressure has built up inside your mouth in the meantime. Unvoiced oral stops with a longer voice onset time are called aspirated stops, and these are the sounds that “pop” if you get too close to a microphone without a pop filter. Following standard notation, we’ll use slightly different symbols for aspirated stops—for example, $p^h$, $t^h$, and $k^h$ (with superscripts) to differentiate them from unaspirated stops $t$, $d$, and $k$. From here on, I’ll also follow the standard practice in linguistics, and instead of using quotation marks around individual sounds, I’ll indicate whether I’m referring to phonemes by enclosing them in forward slashes (for example, /b/, /d/), while allophones will appear inside square brackets (for example, [t], [tʰ]).

Now, notice that a difference in voice onset time is exactly the way I earlier described the distinction between the phonemes /t/ and /d/—sounds that are distinguished in minimal pairs and that cause sudden shifts of meaning (see Figure 4.7). The difference between /t/ and /d/ seems obvious to our ears. And yet we find it hard to notice a similar (and possibly even larger) difference in VOT between the [t] and [tʰ] sounds in Stan and tan. If we become aware of the difference at all, it seems extremely subtle. Why is this? One possible explanation might be that differences at some points in the VOT continuum are inherently easier to hear than distinctions at other points (for example, we might find that the human auditory system had a heightened sensitivity to VOT differences between 10 and 30 ms, but relatively dull perception between 30 and 60 ms). But another possibility is that our perceptual system has become tuned to sound distinctions differently, depending on whether those distinctions are allophonic or phonemic in nature. In other words, maybe what we

voiced Describes a sound that involves vibration of the vocal folds; in an oral stop, the vibration happens just about simultaneously with the release of the articulators (say, within about 20 milliseconds) as it does for “b” in the word ban.

unvoiced (voiceless) Describes a sound that does not involve simultaneous vibration of the vocal folds; in a voiceless stop followed by a vowel, vibration happens only after a lag (say, more than 20 milliseconds).

voice onset time (VOT) The length of time between the point when a stop consonant is released and the point when voicing begins.

phonemic inventory A list of the different phonemes in a language.

aspirated stop An unvoiced oral stop with a long voice onset time and a characteristic puff of air (aspiration) upon its release; an aspirated stop “pops” when you get too close to a microphone without a pop filter. Aspirated stop sounds are indicated with a superscript: $p^h$, $t^h$, and $k^h$.

unaspirated stop An unvoiced oral stop without aspiration, produced with a relatively short VOT.
“hear” isn’t determined only by the objective acoustic differences between sounds, but also by the role that sounds play within the language system.

As it happens, different languages don’t necessarily put phoneme boundaries in the same places; they can differ quite dramatically in how they carve up the sound space into phonemic categories. These distinctions make it possible for us to study whether our perception of sounds is influenced by how a language organizes sound into these categories. Speakers of Thai, for example, distinguish not only between the voiced and unvoiced phonemes /t/ and /d/, but also between the aspirated voiceless phoneme /th/ (as in tan) and its unaspirated version /t/ (as in Stan). What this means is that if you’re speaking Thai, whether or not you aspirate your stops makes a difference to the meaning. Slip up and aspirate a stop by mistake—for example, using /th/ rather than /t/—and you’ve uttered a word that’s different from the one you’d intended.

On the other hand, Mandarin, like English, has only two phonemic categories. But unlike English, Mandarin speakers make a meaningful distinction between voiceless aspirated and unaspirated sounds rather than voiced and voiceless ones. To their ears, the differences between /t/ and /th/ is painfully obvious, corresponding to different phonemes, but they struggle to “hear” the difference between [t] and [d].

Looking across languages, it’s hard to make the case that either the difference between voiced and voiceless sounds or the difference between aspirated and unaspirated sounds is inherently more obvious. Different languages latch on to different distinctions as the basis of their phonemic categories. This becomes all the more apparent when you consider the fact that languages differ even in terms of which dimensions of sound distinction they recruit for phonemic purposes, as we saw in Chapter 3.

For example, in English, whether or not a vowel is stretched out in time is an allophonic matter (Box 4.3). Vowels tend to be longer, for instance, just before voiced sounds than voiceless ones, and can also get stretched out for purely expressive purposes, as in “no waaay!”—note that waaay is still the same word as way. There’s no systematic phonemic distinction between long and short vowels. But in some languages, if you replace a short vowel with a longer one, you’ll have uttered a completely different word (for example, if you lengthen the vowel in the Czech word for Sir, you’ll be addressing someone as cheese). In a similar vein, in Mandarin and various other languages described as “tone languages,” the pitch on a vowel actually signals a phonemic difference. You might have just one sequence of vowels and consonants that will mean up to six or seven different words depending on whether the word is uttered at a high, low, or medium pitch, or whether it swoops upwards in pitch, whether the pitch starts high and falls, or whether the pitch rises and then falls. Needless to say, distinctions like these can be exasperatingly difficult to learn for speakers of languages that don’t use tone phonemically. And, as you also saw in Chapter 3, there’s evidence that different brain processes are involved in using these dimensions of sound, depending on the role they play in the language, with tonal differences on words eliciting more left-
hemisphere activity among Mandarin speakers, but more right-hemisphere activity among English speakers.

I’ve just shown you some examples where other languages have elevated sound distinctions to phonemic status, whereas the same distinctions in English have been relegated to the role of mere sound accessories. The reverse can be true as well. For instance, the English distinction between the liquid sounds /r/ and /l/ is a phonemic one; hence, it matters whether you say rice for Lent or lice for rent. But you’ll probably have noticed that this distinction is a dastardly one for new English language learners who are native speakers of Korean or

**BOX 4.3**

**Vowels**

Unlike consonants, vowels are all made with a relatively unobstructed vocal cavity that allows the air to pass fairly freely through the mouth. Their various sounds are accomplished by shaping the mouth in different ways and varying the placement of the tongue. Interestingly, our perceptual systems tend not to be as categorical when hearing vowels as they are when perceiving consonants, and we’re usually sensitive to even small graded differences among vowels that we’d lump into the same category—though there is evidence that experience with a particular language does have an effect on perception.

Vowels are normally distinguished along the features of **vowel height** (which you can observe by putting a sucker on your tongue while saying different vowels), **vowel backness**, **lip rounding**, and **tenseness**.

English has an unusual number of vowel sounds; it’s not uncommon for languages to get by with a mere five or so. Only a couple of vowels ever occur in English as **diphthongs**, in which the vowel slides into an adjacent glide (as in the words bait boat). Below are the IPA symbols for the English vowel sounds, with examples of how they appear in words in a standard American dialect (note their uneasy relationship to English orthography):

<table>
<thead>
<tr>
<th>IPA</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>beet</td>
</tr>
<tr>
<td>i</td>
<td>bit</td>
</tr>
<tr>
<td>ei</td>
<td>bait</td>
</tr>
<tr>
<td>e</td>
<td>bet</td>
</tr>
<tr>
<td>æ</td>
<td>bat</td>
</tr>
<tr>
<td>ø</td>
<td>the</td>
</tr>
<tr>
<td>aj</td>
<td>bite</td>
</tr>
<tr>
<td>aw</td>
<td>bout</td>
</tr>
</tbody>
</table>

**vowel height** The height of your tongue as you say a vowel; for example, e has more vowel height than a.

**vowel backness** The amount your tongue is retracted toward the back of your mouth when you say a vowel.

**lip rounding** The amount you shape your lips into a circle; for example, your lips are very rounded when you make the sound for w.

**tenseness** A feature of vowels distinguishing “tense” vowels such as those in beet and boot from “lax” vowels such as those in bit and put.

**diphthong** A sound made when the sound for one vowel slides into an adjacent glide in the same syllable, as in the word ouch.

![Figure 4.8](image-url) A vowel chart, a graphic illustration of the features of vowels, including English vowels and vowels found in other languages. When symbols are in pairs, the one to the right is the rounded version. Diphthongs like ei are not marked in this chart but represent transitions between vowels.

The features of the English vowels, along with others that don’t occur in English, can be captured graphically in a vowel chart such as the one in **Figure 4.8**.
Japanese—they are very prone to mixing up these sounds. This is because the difference between the two sounds is an *allophonic* one in Korean and Japanese, and speakers of these languages perceive the difference between the two sounds as much more subtle than do native English speakers.

All of this goes to show that when it comes to how we perceive speech, we aren’t just responding to the actual physical sounds out in the world. The way in which we hear sounds also has a lot to do with the structure our minds impose on sounds of speech. These mental structures can have dramatic effects in perceptually boosting some sound distinctions and minimizing others. We no longer interpret distinctions among sounds as gradual and continuous. This is actually a good thing, because it allows us to ignore many sound differences that aren’t meaningful. For example, your typical English voiced [ba] sound might occur at a VOT of 0 ms, and your typical unvoiced [pʰa] sound might be at 60 ms. But your articulatory system is simply not precise enough to always pronounce sounds at the same VOT (even when you are completely sober); in any given conversation, you may well utter a voiced sound at 15 ms VOT, or an unvoiced sound at 40 ms. But your mind is very good at ignoring this articulatory slippage. What you know about the sound structures of your language imposes sharp boundaries, so you categorize sounds that fall within a single phoneme category—even if they’re different in various ways—as the same, whereas sounds that straddle phoneme category boundaries clearly sound different. This way of perceiving sounds is called *categorical perception*, and it’s quite a handy perceptual strategy.

To get a sense of the usefulness of categorical perception in real life, it’s worth thinking about some of the many examples in which we don’t carve the world up into clear-cut categories. Consider, for example, the objects in Figure 4.9. Which of these objects are cups, and which are bowls? It’s not easy to tell, and you may find yourself disagreeing with some of your classmates about where to draw the line between the two (in fact, that line might readily shift depending on whether these objects are filled with coffee or soup). What’s interesting is that this sort of disagreement is not likely to arise when it comes to consonants that hug the dividing line between two phonemic categories.

Such lack of disagreement is a hallmark of categorical perception, and it’s been amply demonstrated in many experiments. One common way to test for categorical perception is called a *forced choice identification task*. The strategy is to have people listen to many examples of speech sounds and indicate which one of two categories each sound represents (for example, /pa/ versus /ba/). The speech sounds are created in a way that varies the VOT in small increments—for example, participants might hear examples of each of the two sounds at 10-ms increments, all the way from –20 ms to 60 ms. (A *negative* VOT value means that vocal fold vibration begins even before the release of the articulators.)

If people were paying attention to each incremental adjustment in VOT, you’d find that at the extreme ends (i.e., at –20 ms and at 60 ms), there would be tremendous agreement about whether a sound represents a /ba/ or a /pa/, as seen in Figure 4.9. Is it a cup or a bowl? The category boundary isn’t clear, as evident in these images, inspired by a classic experiment by linguist Bill Labov (1972). In contrast, the boundary between different phonemic categories is quite clear for many consonants.

---

**WEB ACTIVITY 4.7**

**Categorical versus continuous perception** In this activity, you’ll listen to sound files that will allow you to compare perception of voiced and unvoiced consonants to the perception of pitch and volume.
in Figure 4.10A. In this hypothetical figure, just about everyone agrees that the sound with the VOT at –20 ms is a /ba/, and the sound with the VOT at 60 ms is a /pa/. But, as also shown in Figure 4.10A, for each step away from –20 ms and closer to 60 ms, you see a few more people calling the sound a /pa/.

But when researchers have looked at people’s responses to forced choice identification tasks, they’ve found a very different picture, more like the graph in Figure 4.10B. People agree pretty much unanimously that the sound is a /ba/ until they get to the 20 ms VOT boundary, at which point the judgments flip abruptly. The upshot of all this is that when you’re processing speech sounds, there’s usually no inner mental argument going on about whether to call a sound /ba/ or /pa/. (The precise VOT boundary that separates voiced from unvoiced sounds can vary slightly, depending on the place of articulation of the sounds.)

What sound distinctions do newborns start with?

Put yourself in the shoes of the newborn, who is encountering speech sounds in all their rich variability for the first time (more or less: some aspects of speech sounds—especially their rhythmic properties—do make it through the uterus wall to the ears of a fetus, but many subtle distinctions among sounds will be encountered for the first time after birth). We’ve seen that adults don’t pay equal attention to all sound distinctions—they pay special attention to those that signal differences between phonemic categories. But we’ve also seen that phoneme categories can vary from language to language, and that sound distinctions that are obvious to one language group may be more elusive to another. Clearly, these distinctions have to be learned to some extent. So what is a newborn baby noticing in sounds? Given that she’s unlikely to have formed categories such as /p/ and /b/, since these categories are somewhat language-specific, does this mean that she’s paying attention to every possible way in which sounds might vary in their pronunciation? Remember that sounds can vary along a number of different dimensions, with incremental variation possible along any of these dimensions. Let’s suppose that babies are perceiving continuously rather than categorically (see Figure 4.9) for any of these sound dimensions. In that case, the sound landscape for babies would be enormously cluttered—where adults cope with several dozen categories of speech sounds, babies might be paying attention to hundreds of potential categories.

It takes some ingenuity to test for categorical perception in newborns. Once again, you can’t give these miniature humans a set of verbal instruc-
tions and get back a verbal response that will tell you whether they are per-
ceiving the difference between certain sounds. You're stuck making do with
behaviors that are within the reach of your average newborn—which, admit-
tedly, are not a lot. Faced with a newborn whose behavioral repertoire seems
limited to sleeping, crying, sucking, and recycling body wastes, a researcher
might be forgiven for feeling discouraged. It turns out, though, that one of
these behaviors—sucking—can, in the right hands, provide some insight into
the infant’s perceptual processes. Babies suck to feed, but they also suck for
comfort, and if they happen to have something in their mouths at the time,
they suck when they get excited. And, as may be true for all of us, they tend
to get excited at a bit of novelty.

By piecing these observations together, Peter Eimas and his colleagues
(1971), pioneers in the study of infant speech perception, were able to design
an experimental paradigm that allows researchers to figure out which sounds
babies are perceiving as the same, and which they’re perceiving as different.
The basic premise goes like this: If babies are sucking on a pacifier while hear-
ing speech sounds, they’ll tend to suck vigorously every time they hear a new
sound. But if they hear the same sound for a long period of time, they become
bored and suck with less enthusiasm. This means that a researcher can cleverly
rig up a pacifier to a device that measures rate of sucking, and play Sound A
(say, [pa]) over and over until the baby shows signs of boredom (that is, the
baby’s sucking slows down). Once this happens, the researcher can then play
Sound B (say, [pʰa]). If the baby’s sucking rate picks up, this suggests the baby
has perceived Sound B as a different sound. If it doesn’t, it provides a clue that
the baby, blasé about the new sound, hasn’t perceived it as being any different
from the first one (see Method 4.2 for details of this approach). When it comes
to testing for categorical perception then, if babies perceive speech sounds cat-
ergically, they should be oblivious to differences between certain sounds but
acutely sensitive to differences between other sounds that fall on different sides
of a critical boundary. On the other hand, if they’re perceiving continuously,
then they should always hear Sound B as different, and should increase their
sucking just about any time Sound B is introduced.

If we look at how babies perceive VOT, the experiments show clear evidence
of categorical perception in newborns, so it appears that the youngest humans
don’t treat all sound distinctions in the same way. Their rate of sucking goes
up when two sounds straddle a VOT boundary of about 25 ms, but otherwise
they seem oblivious to differences in VOT. This boundary is very similar to the
adult dividing line for English voiced and voiceless sounds. What this means is
that the sound landscape comes pre-carved to some extent; upon birth, babies
aren’t faced with the massive task of considering every possible difference in
sound as being potentially meaningful when it comes to signaling differences
between phonemic categories. Some sound distinctions are more privileged
than others right off the bat.

When researchers first discovered that babies emerge from the womb with
certain pre-set boundaries that happen to line up with the VOT boundaries
distinguishing English voiced and voiceless sounds, this generated some ex-
cited speculation. Some researchers suggested that children come innately
equipped with a set of inborn phonetic categories that are commonly used by
languages. But this line of thinking quickly ran into a wall. First of all, Patricia
Kuhl and James Miller (1971) devised a clever experiment to study the percep-
tion of consonants by chinchillas—which, while adorable, are not known for
their linguistic skills, and certainly don’t ever produce speech, so it’s doubtful
that they would be born innately prepared for it. Kuhl and Miller found that
these small, furry mammals also perceived consonants categorically, along a VOT boundary very similar to the one found for humans (see Box 4.4). This result has since been replicated in some of our closer relatives, such as macaque monkeys, and in much more distant animal relatives such as birds.

There’s a second problem with the notion that categorical perception in human newborns reflects innate preparation for linguistic sounds: it turns out that many non-speech sounds are perceived categorically as well—not just by humans but also by animals that are very distant from us on the evolutionary family tree, such as crickets and frogs. So it looks as if the process of amplifying some sound distinctions while minimizing others is a very general property of the auditory system across species. Though it has a certain usefulness for perceiving speech, it doesn’t seem to be intrinsically related to speech.

An especially telling demonstration of the parallels in perception of speech and non-speech sounds comes from experiments that use non-speech sounds to mimic some of the properties of human speech sounds. Remember that VOT

**Method 4.2**

**High-amplitude sucking**

The high-amplitude sucking method allows researchers to peer into the minds of babies who, due to their tender age, understandably have a limited repertoire of behaviors. It’s based on the premise that infants will naturally suck on objects in their mouths when they are excited by hearing a new sound. Throughout the experiment, the baby participant sucks on a pacifier that contains electrical circuitry that measures the pressure of each sucking motion so that the rate of sucking can be constantly tracked. The pacifier is held in place by an assistant, who wears headphones and listens to music to block out the experimental sounds, so there’s no possibility of sending any inadvertent signals to the baby.

Infants tend to suck with gusto when they hear a new, interesting sound anyway, but in order to get the strongest connection possible between a new stimulus sound and this sucking behavior, researchers have an initial conditioning phase built into the session. During this phase, a new sound is played every time the baby sucks at a certain rate: no vigorous sucking, no terrific new sound. Babies quickly learn to suck to hear the stimulus.

Once the baby has been trained to suck to hear new sounds, researchers play the first of a pair of stimulus sounds—let’s say a [pa] sound—over and over. Once the baby’s interest lags, the sucking rate goes down. This is an example of habituation. When the baby’s sucking rate dips below a criterion level that’s been previously established, the second sound of the pair is played, and the dependent measure is the sucking rate of the baby after the presentation of this new sound. If the baby begins to suck eagerly again, it’s a good sign that she perceives the second sound as different from the first.

This method can be adapted to measure which of two kinds of stimuli infants prefer. For example, if you wanted to know whether infants would rather listen to their own language or an exotic tongue, you could set up your study in a way that trains babies to suck slowly or not at all in order to hear one language, and to suck hard and fast in order to hear the other (being sure to counterbalance your experimental design to make sure that an equal number of babies are trained to suck slowly versus quickly for the native language).

As you might imagine, when working with extremely new babies, there can be quite a lot of data loss. Often babies are too tired, too hungry, or just too ornery to pay much attention to the experimental stimuli, so a large number of participants must be recruited for this method. The method works quite well for infants up to about 4 months of age, after which point the pastime of sucking begins to lose some of its appeal for infants, and they’re less likely to keep at it for any length of time. Luckily, at around this age the head-turn preference paradigm becomes an option for testing babies’ perception of speech.

**habituation**

Decrease in responsiveness to a stimulus upon repeated exposure to the stimulus.
Learning Sound Patterns

is a measure of the time between the release of the articulators and the beginning of voicing (that is, vibration of the vocal folds). A slightly more abstract way of looking at it is that the perception of VOT is about perceiving the relative timing of two distinct events. This scenario can easily be recreated with non-

**BOX 4.4**

**Categorical perception in chinchillas**

How can you study speech perception in animals? As they do with babies, scientists have to find a way to leverage behaviors that come naturally to animals, and incorporate these behaviors into their experiments. Speech scientists Patricia Kuhl and James Miller (1975) took advantage of the fact that in many lab experiments, animals have shown that they can readily link different stimuli with different events, and that they can also learn to produce different responses to different stimuli in order to earn a reward.

In Kuhl and Miller’s study, chinchillas (Figure 4.11A) heard various speech sounds as they licked a drinking tube to get dribbles of water. When the syllable /da/ was played (with a VOT of 0 ms), it was soon followed by a mild electric shock. As would you or I, the chinchillas quickly learned to run to the other side of the cage when they heard this sound. On the other hand, when the syllable /ta/ came on (with a VOT of 80 ms), there was no electric shock, and if the chinchillas stayed put and continued to drink, they were rewarded by having the water valve open to allow a stronger flow of water. In this way, the researchers encouraged the chinchillas to link two very different events to the different phonemic categories, a distinction the chinchillas were able to make.

The next step was to systematically tweak the voice onset times of the speech stimuli in order to see how well the chinchillas were able to detect differences between the two sounds at different points along the VOT continuum. Even though the animals had only heard examples of sounds at the far ends of the VOT spectrum, they showed the same tendency as human babies and adults do—that is, to sort sounds into clear-cut boundaries (see Figure 4.10)—and the sharp boundary between categories occurred at almost the same VOT as was found for humans (33.5 ms versus 35.2 ms; Figure 4.11B).

**Figure 4.11** (A) A chinchilla; these animals are rodents about the size of a squirrel. They are a good choice for auditory studies because the chinchilla’s range of hearing (20–30 kHz) is close to that of humans. (B) Results from Kuhl and Miller’s categorical perception experiment, comparing results from the animals and human adults. The graph shows the mean percentage of trials in which the stimulus was treated as an instance of the syllable /da/. For humans, this involved asking the subjects whether they’d heard a /da/ or /ta/ sound; for chinchillas, it involved seeing whether the animals fled to the other side of the cage or stayed to drink water. (After Kuhl and Miller, 1975.)

![Chinchilla](image-url)
speech stimuli, simply by putting together two distinct sounds and playing around with their relative timing.

Researcher David Pisoni (1977) created a set of stimuli by using two distinct tones and varying the number of milliseconds that elapsed between the onsets of the two tones, much as was done in previous VOT experiments—we can call this “tone onset time,” or TOT. He then tested to see whether there was a certain window across which people would be especially sensitive to TOT differences. For instance, people might hear two stimuli, Stimulus A being two tones whose onsets were separated by 20 ms, and Stimulus B being two tones separated by 30 ms. The people would then have to judge whether a third stimulus (in which the two tones were separated by, say, 30 ms) was the same as Stimulus A or Stimulus B. The idea behind this task, known as an ABX discrimination task, is that if people can readily perceive the difference between the two sound pairs, they’ll be reliable at identifying whether the third sound pair is identical to the first or second. On the other hand, if they don’t perceive the difference between them, then they’ll be randomly guessing as to the identity of the third sound pair.

What Pisoni found was that people were especially good at distinguishing between stimuli right around a TOT of 20 ms. For example, the above pair of stimuli, sound pairs with TOTs of 20 ms and 30 ms, would be perceived as distinct by many of the subjects. But if people heard a pair of stimuli with sounds separated by 40 ms and 50 ms, they were much less likely to perceive them as different. The same was true for a pair of simultaneously produced sounds (0 ms TOT) and a sound pair 10 ms apart. In other words, the TOT boundary for optimal perception of differences was strikingly similar to the boundary for voice onset time of speech sounds. Pisoni suggested that differences at about the 20 ms boundary for both speech and non-speech sounds are easy to notice because this is the point at which the auditory system is able to detect that two events occurred at different times. If the time between two events is any shorter, it becomes hard to perceive that they didn’t occur at the same time. The limits of the auditory system make the 20 ms mark a point at which stimuli naturally divide up into categories of simultaneous versus non-simultaneous pairs of sound events.

Clearly, the overall evidence from categorical perception scores no points for the hypothesis that babies come preinstalled with probable speech categories. Instead, it supports the notion that a language like English is being opportunistic about where it carves phonemic categories—it appears to be shaping itself to take advantage of natural perceptual biases of the auditory system.

Still, not all languages take advantage of the natural places to carve up phonemic categories that the auditory system so conveniently offers up. As we saw, languages like Mandarin opt not to distinguish between phonemes at the “natural” boundaries, placing phonemic boundaries elsewhere instead. Since babies are obviously able to grow into Mandarin-speaking adults, there must be enough flexibility in their perceptual systems to adapt to the categories as defined by their particular language. What changes in the perceptual life of an infant as she digests the sounds of the language around her?

Quite a bit of research has shown that babies start off noticing a large number of distinctions among sounds, regardless of whether the languages they’ll eventually speak make use of them to mark phonemic distinctions. For example, all babies, regardless of their native languages, start off treating voiced and unvoiced sounds as different, and the same goes for aspirated versus unaspirated sounds. As they learn the sound inventory of their own language, part of their job is to learn which variations in sounds are of a deep, meaning-
changing kind, and which ones are like wardrobe options. Eventually, Mandarin-hearing babies will figure out that there’s no need to separate voiced and voiceless sounds into different categories, and they will downgrade this sound difference in their auditory attention. (Here’s a workable analogy to this attentional downgrading: presumably, you’ve learned which visual cues give you good information about the identity of a person, and which ones don’t, so you pay more attention to those strongly identifying cues. So, you might remember that you ran into your co-worker at the post office, but have no idea what she was wearing at the time.) Unlike the Mandarin-hearing babies, who “ignore” voicing, English-hearing babies will learn to “ignore” the difference between aspirated and unaspirated sounds, while Thai-hearing babies will grow up maintaining a keen interest in both of these distinctions.

It can be a bit humbling to learn that days-old babies are good at perceiving sound differences that you strain to be aware of. There’s a large body of research (pioneered by Janet Werker and Richard Tees, 1984) that now documents the sounds that newborns tune in to, regardless of the language their parents speak. Unlike many of you, these tiny bundles of joy can easily cope with exotic sound distinctions, including these: the subtle differences among Hindi stops (for instance, the difference between a “regular” English-style /t/ sound and one made by slightly curving the tip of your tongue back as you make the sound); Czech fricatives (for instance, the difference between the last consonant in beige and the unique fricative sound in Dvorak); and whether a vowel has a nasal coloring to it, a distinctive feature in French, important for distinguishing among vowels that shift meaning. At some point toward the end of their first year, babies show evidence of having reorganized their perception of sounds. Like adults, they begin to confer special status on those distinctions that sort sounds into separate phonemic categories of the language they’re learning.

4.4 Learning How Sounds Pattern

The distribution of allophones

In the previous section, we saw that babies start by treating many sound distinctions as potentially phonemic, but then tune their perception in some way that dampens the differences between sounds that are non-phonemic in the language they’re busy learning. So, a Mandarin-hearing baby will start out being able to easily distinguish between voiced versus unvoiced sounds, but will eventually learn to ignore this difference, since it’s not a phonemic one.

But this raises a question: How do infants learn which sounds are phonemic, and therefore, which differences are important, and which can be safely ignored? You and I know that voicing is a distinctive feature partly because we recognize that bat and pat are different words with different meanings. But remember that babies are beginning to sort out which sound differences are distinctive as early as 6 months of age, at a time when they know the meanings of very few words (a topic we’ll take up in Chapter 5). If infants don’t know what bat and pat mean, or even that bat and pat mean different things, how can they possibly figure out that voicing (but not aspiration) is a distinctive feature in English?

As it happens, quite aside from their different roles in signaling meaning differences, phonemes and allophones pattern quite differently in language. And, since babies seem to be very good at noticing statistical patterns in the
language they’re hearing, such differences might provide a useful clue in helping them sort out the phonemic status of various sounds.

Let’s take aspiration as a test case. Monolingual English speakers have a hard time consciously distinguishing a [p] from a [pʰ], especially in tasks that focus on sorting these sounds into categories. Yet the speech patterns of these same speakers show that they must be able to hear the difference between them at some level, because they produce them in entirely different sound contexts. Aspirated sounds get produced at the beginnings of words or at the beginnings of stressed syllables—for example, the underlined sounds in words like *pit, cat, PAbllum, CAsle, baTTAlion, comPAssion*. Unaspirated sounds get pronounced when they follow another consonant in the same syllable—like the underlined sounds in words like *spit, stairs, scream*, and *schtick*—or when they’re at the beginning of an unaccented syllable, as in *CAnter, AMputate, eLLIptical*. The /p/ sound would come off as aspirated in *comPUter*, but not in *COMputational*. In other words, aspirated and unaspirated sounds aren’t sprinkled throughout the English language randomly; they follow a systematic pattern that speakers must have somehow unconsciously noticed and reproduced even though they think they can’t hear the difference between them. (In case you’re wondering, it’s perfectly possible from a purely articulatory standpoint to produce a word-initial voiceless sound without aspiration, even when it’s at the beginning of a stressed syllable—speakers of many other languages do it all the time).

I’ve talked about how allophonic distinctions are like wardrobe changes for the same sound. Well, just as you wouldn’t wear your fishnet stockings to a funeral, or your sneakers to a formal gala, allophones tend to be restricted to certain environments—there are “rules” about which allophones can turn up in which places. And just like “rules” about wardrobe choices, they are to some extent a bit arbitrary, and based on the conventions of a language. When two allophones are relegated to completely separate, non-overlapping linguistic environments, they’re said to be in complementary distribution. In fact, showing up in different linguistic environments from each other is a defining feature of allophones, along with the fact that they don’t signal a change in meaning. This means that whenever a sound distinction is allophonic, rather than phonemic, it should be possible to predict which of the two sound variants will turn up where.

We can see how this works with an example other than aspiration. In English, certain differences among sounds involving the place of articulation are distinctive, but others are not. For instance, sounds produced by placing the tip of the tongue against the alveolar ridge are phonemically distinct from sounds produced by the back of the tongue (for example, voiceless /t/ versus /k/ and voiced /d/ versus /g/). One upshot of this is that, taken completely outside of their semantic context (that is, the context of meanings), it’s impossible to predict whether the alveolar sounds or the back (velar) sounds will fill in the blanks below (The sounds are depicted in International Phonetic Alphabet, or IPA symbols, as in Figure 4.6 and Box 4.3.)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>__ep (as in the word <em>ape</em>)</td>
<td>__ Il (as in <em>fill</em>)</td>
</tr>
<tr>
<td>__ an (as in <em>lawn</em>)</td>
<td>__ owp (as in <em>nope</em>)</td>
</tr>
<tr>
<td>__ il (as in <em>feel</em>)</td>
<td>__ un (as in <em>swoon</em>)</td>
</tr>
<tr>
<td>__ ân (as in <em>fan</em>)</td>
<td>__ ik (as in <em>sick</em>)</td>
</tr>
</tbody>
</table>

In all of these blanks, you can insert either an alveolar or a velar (back) sound, often of either the voiced or voiceless variety: *tape/cape; Don/gone; teal/keel; tan/can; dill/gill/till/kill; tape/cope; tune/coon/dune/goon; tick/kick.*

**complementary distribution** Separation of two allophones into completely different, non-overlapping linguistic environments.
But not all distinctions that involve place of articulation are phonemic in English. For instance, there’s an allophonic distinction between the velar stop [k] and a palatal stop [c], which is made a little farther forward than [k], up against the palate. In each of the same linguistic environments you just saw, only one of either the velar or palatal sounds tends to show up—that is, the sounds are in complementary distribution (see Box 4.5). Their distribution is shown below (ignore standard English orthography, and look instead at how the sounds are represented as IPA symbols):

<table>
<thead>
<tr>
<th>Nasalized</th>
<th>Non-nasalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>bit (beet)</td>
<td>bɛn (been)</td>
</tr>
<tr>
<td>lowd (load)</td>
<td>lɛwn (loan)</td>
</tr>
<tr>
<td>braɛ (braɛ)</td>
<td>bɹɛnd (brand)</td>
</tr>
<tr>
<td>bɛd (bed)</td>
<td>bɛnd (bend)</td>
</tr>
<tr>
<td>bat (but)</td>
<td>bɛt (bent)</td>
</tr>
<tr>
<td>brud (brood)</td>
<td>brʌm (broom)</td>
</tr>
<tr>
<td>sɪk (sick)</td>
<td>sɪŋ (sing)</td>
</tr>
</tbody>
</table>

Finnish/English
Vowel length marks a phonemic distinction in Finnish, but in most English dialects, differences in vowel length mark different allophones of the same phoneme. In English, longer vowels appear before voiced sounds in the same syllable, while shorter vowels appear before unvoiced sounds:

<table>
<thead>
<tr>
<th>Short</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>lɪt (lit)</td>
<td>lɪ:d (lied)</td>
</tr>
<tr>
<td>rowp (rope)</td>
<td>row:b (robe)</td>
</tr>
<tr>
<td>sɛt (set)</td>
<td>sɛ:d (said)</td>
</tr>
<tr>
<td>mæt (mat)</td>
<td>mæ:d (mad)</td>
</tr>
<tr>
<td>fɛt (fate)</td>
<td>fɛt:d (fade)</td>
</tr>
</tbody>
</table>

English/Spanish
The voiced alveolar stop consonant /d/ and its closest corresponding fricative /ð/ (as in the word then) are different phonemes in English; hence the existence of minimal pairs such as den and then. In most dialects of Spanish, however, these two sounds are allophones, and therefore in complementary distribution. The fricative appears after a vowel in Spanish:

<table>
<thead>
<tr>
<th>Stop</th>
<th>Fricative</th>
</tr>
</thead>
<tbody>
<tr>
<td>dɔnde (where)</td>
<td>usteð (you)</td>
</tr>
<tr>
<td>dolor (pain)</td>
<td>comiða (food)</td>
</tr>
<tr>
<td>durar (to last)</td>
<td>laðo (side)</td>
</tr>
<tr>
<td>andar (to walk)</td>
<td>caða (each)</td>
</tr>
<tr>
<td>falda (skirt)</td>
<td>oðio (hatred)</td>
</tr>
</tbody>
</table>

But not all distinctions that involve place of articulation are phonemic in English. For instance, there’s an allophonic distinction between the velar stop [k] and a palatal stop [c], which is made a little farther forward than [k], up against the palate. In each of the same linguistic environments you just saw, only one of either the velar or palatal sounds tends to show up—that is, the sounds are in complementary distribution (see Box 4.5). Their distribution is shown below (ignore standard English orthography, and look instead at how the sounds are represented as IPA symbols):

| c ɛɪp (cape) | c ɪl (kill) | c ɪl (keel) | k un (coon) |
| k ʌn (con)   | k ʌwp (cope) | k ʌn (can) | c ɪk (kick) |
If you were to look at many more words of English, a clear generalization would emerge: [c] is allowed whenever it comes before vowels such as /e/, /ɪ/ or /i/; [k] on the other hand shows up in front of sounds like /ɑ/, /o/ or /u/. You might then also notice that /əl, /ɪl, and /i/ have something in common—they’re produced at the front of the mouth, while /ɑl, /ɪl, or /i/ are produced at the back of the mouth. It’s probably no accident that the palatal sound [c]—which is produced farther forward in the mouth—is the sound that appears with the front vowels rather than the other way around. It’s often the case that an allophonic variant will resemble adjacent sounds in some way, following a natural process called assimilation. It’s important to realize, though, that while the rule that determines the distribution of [k] and [c] is a fairly natural one, given that the stops often morph to resemble their adjacent vowels, there’s nothing inevitable about it. In Turkish, for example, /k/ and /c/ are separate phonemes, and each stolidly maintains its shape regardless of which vowel it’s standing next to.

**From patterns of distribution to phonemic categories**

We know from the work on word segmentation that babies are extremely good at using statistical information about sound patterns to guess at likely word boundaries. It makes sense to ask, then, whether babies might also be able to use information about how sounds like [k] and [c] pattern in order to figure out which of the sounds in their language are phonemic. By “noticing” (implicitly, rather than in any conscious way) that these sounds are in complementary distribution, babies born to English-speaking parents might conclude that they’re allophones and stop perceiving the sounds categorically as representing two phonemes. That is, they would start to treat [k] and [c] as variants of one sound. Babies in Turkish-speaking households, on the other hand, would have no reason to collapse the two sounds into one category, so they would stay highly tuned to the distinction between /k/ and /c/.

Experiments by Katherine White and her colleagues (2008) suggest that babies might be using distributional evidence along these lines to categorize sounds as “same” versus “different.” To test this, they used the trick of devising an artificial language with certain statistical regularities and checked to see what babies gleaned from these patterns. In this language, babies heard the following set of two-word sequences, repeated in random order:

```
na bevi rot pevi na zuma rot zuma na suma rot suma
na bogu rot pugu na zobi rot zobi na sibi rot sibi
na dula rot tula na veda rot veda na feda rot feda
na dizu rot tizu na vadu rot vadu na fadu rot fadu
```

Do you see the pattern? The babies did. When it comes to the word-initial stops (that is, “b,” “p,” “t,” “d”), whether they are voiced or not depends on the preceding words—if the preceding word ends in a voiceless sound (rot), then the stop is also voiceless, assimilating to the previous sound. But if the preceding word ends in a voiced sound (that is, na—remember, all vowels are by default voiced), then the stop is also voiced. When you look at the fricatives, though, either voiced or voiceless fricatives (“z,” “s,” “ʃ,” “v”) can appear regardless of whether the last sound of the preceding word is voiced or voiceless. In other words, stops are in complementary distribution, but fricatives are not.

Now, based on these patterns, would you think that bevi and pevi are different words, or just different ways of pronouncing the same word? Given that “b” and “p” are in complementary distribution, and therefore that they are likely allophones of the same phoneme, switching between “b” and “p” prob-
ably doesn’t change the meaning of the word—it’s just that it’s pronounced one way after *na* and a slightly different way after *rot*, so you can count entirely on the distributional rules of the language to figure out which variant it should be in that context. What about *zuma* and *suma*? Here, the voiced and voiceless sounds “z” and “s” aren’t constrained by some aspect of the phonetic context, so it would make sense to assume that they’re separate phonemes. Which of course means that the words *zuma* and *suma* are probably minimal pairs, each with a different meaning.

Katherine White and colleagues found (using the head-turn preference paradigm) that at 8.5 months of age, the babies had caught on to the fact that stops and fricatives involved different patterns of distribution when it comes to voicing—they listened longer to “legal” sequences involving new words beginning with a stop (for example, *rot poli, na boli, rot poli, na boli*) than they did to sequences beginning with a fricative (for example, *rot zadu, rot sadu, rot zadu, rot sadu*). This may be because the babies were able to predict the voicing of the word-initial stop—but not the fricative—based on the previous word, so the words involving stops may have felt a bit more familiar. So, by 8.5 months, babies were able to tune in to the fact that there was a special relationship between a stop and the preceding sound, but that this predictive relationship was absent for fricatives. By 12 months of age, they seemed to understand that this relationship had something to do with whether word units that differed just in the voicing of their first sounds should be treated as “same” or “different” units. At this age (but not at the younger age), the babies also showed a difference in their listening times for sequences of words with stops versus words with fricatives, even when they appeared *without* the preceding word (that is, *poli, boli, poli, boli* versus *zadu, sadu, zadu, sadu*). This makes sense if they were treating *poli* and *boli* as variants of the same word but thinking of *zadu* and *sadu* as different words.

Babies aren’t ones to waste information—if they find a pattern, they’re likely to put it to good use. As we’ve seen, distributional sound patterns, which link specific sounds to the phonetic contexts where they can occur, are very handy for inferring which sounds are distinct phonemes rather than variants of a single phoneme. They can also provide some clues about where the boundaries are between words. This is because sometimes whether you pronounce one allophone or another depends upon whether it’s at the beginning or end of a word. Compare *night rate* and *nitrate*, for example. Both are made from the same sequence of phonemes strung together, and the only difference is whether there’s a word boundary between the /t/ and /r/ sounds. In normal speech, there would be no pause at all between the words. But this word boundary has subtle phonetic consequences nevertheless: If you say *night rate* at a normal conversational pace, the /t/ sound in *night* is unaspirated, and also unreleased (notice that once the tongue meets the alveolar ridge, it kind of stays there as you slide into the /r/ sound). On the other hand, when you say *nitrate*, the first /t/ sound is aspirated and audibly released, and what’s more, the following /r/ sound, which is usually voiced, becomes voiceless by virtue of assimilating to the /t/ sound before it. If babies have noticed that the sounds /t/ and /r/ take on a slightly different shape depending on whether there’s a word boundary between them or not, this might help them make better guesses about whether *night rates* and *nitrates* form one word unit or two.

Peter Jusczyk and his colleagues (1999) showed that by 10.5 months of age, babies who heard *night rates* during the familiarization phase of a study were later able to distinguish this phrase from the nearly identical *nitrates* (based on the result that during the test phase, they listened longer to the familiar *night rates* than to the novel word *nitrates*). This shows that the babies were tuning in
to the very subtle differences in sounds between these two sound sequences, and probably weren’t hearing nitrates as the same word as night rates.

It’s not hard to see how this kind of information about likely word boundaries could come in very handy in helping babies to avoid mistakes in slicing up the speech stream. A subsequent study by Mattys and Jusczyk (2001) showed that even at 8.5 months, babies didn’t treat the word dice as familiar if they’d previously heard a sentence like The city truck cleared ice and sand from the sidewalk—here, the sound sequence d-ice appears, but with a word boundary after the /d/ sound, which affects how the sequence is pronounced. The babies were not fooled into thinking they’d heard the word dice. But the youngsters did seem to recognize the word dice if instead they previously heard the sentence Many dealers throw dice with one hand.

If anything, this chapter ought to have cured you of any tendency to underestimate the intelligence of babies, or to believe that they’re not paying attention to what you say. Clearly, a spectacular amount of learning goes on behind the innocent eyes of infants in their first year. Especially over the second half of that first year, we see piles of evidence that babies’ knowledge of the sound system of their language is undergoing dramatic learning and perceptual reorganization. Before uttering their first words, babies have become competent at chopping up the continuous flow of speech into word-like units, figuring out which sound distinctions define sound categories for their particular language, hearing how subtle differences in sound might be related to the phonetic context in which those sounds appear, and leveraging that information in a number of useful ways.

Of course, we haven’t said anything at all yet about what the little tykes do with this vast knowledge of their language’s sounds. Mapping these sounds onto meanings is a whole other task, one that we’ll take up in the next chapter.

### DIGGING DEEPER

**Statistics, yes, but what kind of statistics?**

We’re not used to thinking of infants as having great powers of statistical analysis, and yet the scientific work on infant speech perception tells a pretty convincing story that babies zero in on exquisitely detailed statistical regularities from the very beginning stages of learning their language. In fact, this attention to statistical detail seems to be the bedrock on which later language learning can be built.

In many ways, the scientific story is just beginning. Though we have good evidence that babies (along with other animals) pick up on certain kinds of statistical regularities, there are many things we still don’t know.

When you stop to think about it, the statistical regularities that could be entertained by babies come in a great number of varieties and flavors. Do infants focus on some more than others? Does the nature of the patterns they look for change over time? Are the types of patterns they can track different from the patterns that other species of animals are able to track? Do babies notice different kinds of patterns in language than they do in other perceptual domains? And do they have some inborn sense of just which kinds of statistical patterns might be the most useful for the learning of a language?

As you can see, we’ve just begun to scratch the surface. Without digging in and conducting a very large number of studies looking at what may seem like very small details, we can’t answer the big questions, such as whether statistical learning has an innate component, or whether humans do it differently from animals.

Let’s begin to make all this a bit more concrete. We’ve seen that babies as young as 8 months of age can track the transitional probabilities (TPs; see p. 117) in a language—that is, they infer that for any two syllables, if the first syllable provides a strong cue to the identity of the second syllable, then those two syllables are quite likely to be grouped together in a word. So, the two syllables in blender are good candidates for a word unit because, given the first syllable blen, there’s a fairly
high likelihood that you’ll hear the second syllable der. This
of course assumes that babies are computing probabilities in
a particular direction, from left to right. But in principle, it’s
also perfectly reasonable to ask whether, given the second
syllable der, there’s a high likelihood that it will have been
preceded by the first syllable blen. In other words, babies
could also be computing backward transitional probabilities.

It seems a bit odd to think about tracking statistical
relationships this way, because we’re so used to thinking
of language in a left-to-right direction. But it turns out that
backward TPs are just as useful for figuring out whether or
not two syllables are part of the same word. So, computing
them in either direction is likely to be helpful in terms of
identifying the statistical peaks and valleys that provide cues
about likely word boundaries.

In fact, there are some cases where backward TPs might
be more informative in identifying certain regularities,
especially when it comes to thinking about grammatical
relationships between words. Suppose, for instance, you are
a baby trying to figure out whether the word bottle is a noun
or a verb. A really strong cue for noun-hood is that nouns
tend to be preceded by articles such as a or the. In other
words, transitional probabilities can provide some good cues,
but in this case, they need to be of the backward variety—
that is, it would be very helpful to have noticed that, given
the word bottle, there was an extremely high likelihood
that the preceding word was the. Looking only at forward
TPs would be less helpful, because given the word the, the
likelihood of its being followed by bottle is fairly low.

Backward and forward TPs often tend to correlate with
each other in natural languages. But it’s possible to carefully
set up experimental stimuli from an artificial language or a
completely unfamiliar language such that either the forward
or backward TPs are more informative than the other, in
order to test whether babies are sensitive to both sources of
information. Using this strategy, Bruna Pelucchi et al. (2009)
found that at 8 months of age, babies could indeed track
backward TPs as a way to extract words from the speech
stream of an unfamiliar language. So infants don’t seem to
be constrained to computing statistical regularities in one
direction only.

Another issue that cries out for exploration is how much
phonetic detail is statistically tracked. For instance, should
stress be marked on syllable units over which statistics are
tracked? So far, we’ve been implicitly assuming that stress
as a cue to word segmentation is separate from TPs. This
means that in order to compute TPs, the test in the words
CONtest and deTEST would be counted as the same syllable.
If you think back to our discussion of stress in Section 4.1,
I pointed out that in English, the majority of words have a
trochaic stress pattern, with stress on the first syllable, as in
CONtest. I talked about how, once babies have figured out
this generalization, it could be useful to them in segmenting
new sequences of sounds they hadn’t heard before: when
in doubt, put the word boundary to the left of the stressed
syllable. Of course, babies could only notice that most words
in English are trochaic once they’d accumulated enough
English words! If you think of stress and TPs as separate in
this way, it makes sense to predict that babies would rely on
stress as a cue only sometime after they were able to rely on
using TPs to segment the speech stream. This is because they
would first use the TPs as a way to amass a large enough
collection of words over which to generalize about stress.

There’s some evidence to support the idea that statistical
cues to word segmentation are used some time before stress
can be applied: Thiessen and Saffran (2003) showed that 6-
and 7-month-olds could use TPs to segment new words from
an artificial language but that they didn’t show any tendency
to fall back on a trochaic segmentation bias; 9-month-olds,
on the other hand, put more stock in the stress cues when
they confronted with the bare statistical cues.

Another way of looking at this is that what changes
with a baby’s age is that by 9 months, babies have learned
to incorporate stress as part of the information that goes
into computing TPs. The idea would be that the 9-month-
olds were treating the test in CONtest and deTEST as two
different syllables. This has the effect of turbocharging TPs:
Curtin et al. (2005) analyzed all the pairs of syllables in a
body of English speech directed at babies, and they found
that including stress in the calculation of TPs created an
even wider separation between transitional probabilities
for within-word syllable pairs than for across-word syllable
pairs. In other words, statistical cues became quite a bit more
reliable once the information about stress was folded in. The
developmental change, then, may be that over time, babies
incorporate details about sounds to the extent that they
figure out that doing so will enhance the statistical cues. This
would make them sophisticated statisticians indeed.

It would almost seem as if eventually, if there are
statistical regularities to be found, babies find them. This
might logically lead you to think that babies are built to
be able to pick up on any kind of statistical regularity.
But so far, all of the variations on statistical cues we’ve
looked at involve what are really fairly minor tweaks of
the original TPs as stated by Saffran et al. (1996). If you get
wildly imaginative about the different possible statistical
relationships between sounds, you can cook up some truly
unusual generalizations. For instance, imagine a language
in which the last sound of a word is always /m/ if the word
happens to begin with a /k/ sound, and is always /s/ if the
word begins with the vowel /o/. It’s perfectly possible to
create an artificial language in which this generalization is
absolutely regular—for example, kabidestim, kum, kendom,
obalid, otis, ojadiguntines. But it’s a type of generalization
that is extremely unlikely to show up in a natural language,
despite the seemingly unlimited diversity of languages. Real
languages tend to stick to regularities that are stated in terms
of adjacent or near-adjacent elements.
Once you start looking at sound regularities across many different languages, a number of constraints and typical patterns start to emerge. For example, think of the generalization that determines whether voiceless sounds in English will be aspirated or unaspirated. Notice that it applies to all of the voiceless sounds of English, not just one or two of them. It would be a bit weird for a language to have [pʰam] and [kʰam] but [tam], with an unaspirated [t] in this position. That is, sound regularities usually apply to natural classes of sounds—that is, groups of sounds that are very similar to one another in phonetic space and that share quite a few articulatory features. For aspiration in English, you can make a broad generalization that voiceless stops become aspirated in certain linguistic contexts, without having to specify individual sounds.

It’s also true that allophones of a single phoneme tend to have a lot in common phonetically—so, think of [p] and [pʰ], but also of the liquid sounds [t] and [l], which are allophones in Japanese and Korean. This means that it would be strange for two completely different sounds—say, [ɾ] and [f]—to be in complementary distribution with each other, even though from a strictly mathematical point of view, there’s nothing to prevent it.

So, there seem to be some constraints on which sounds or groups of sounds are typically the targets of sound-based generalizations. What’s more, there are also some constraints on the types of contexts or neighboring sounds that tend to determine which variant of a sound will end up being pronounced. As we’ve noted before, the plural marker –s as in dogs and cats is actually pronounced differently in these two words. Attach it to cat, and you utter the voiceless fricative [s], but tag it onto dog, and you pronounce its voiced sibling [z]. And if you start paying attention to all regular plural forms, you’ll find that the voiced fricative shows up whenever the immediately preceding sound is voiced, and that it is voiceless whenever it comes on the heels of a voiceless sound. (Notice how it’s pronounced in dogs, docks, cats, cabs, caps, cabs, and so on.) It’s probably not sheer coincidence that the plural marker is affected by an adjacent sound, rather than another sound two syllables over. Nor is it likely to be a coincidence that the feature that undergoes the change in the plural marker—that is, voicing—is also the feature than defines the classes of relevant preceding sounds. Sound patterns like these, in which one feature of a sound bleeds over onto a neighboring one, are extremely common across languages. What would be less common is a pattern in which, say, the plural sound [s] became the stop [t] whenever it followed a voiceless sound, but a fricative whenever it followed a voiced sound.

After surveying the world’s languages, then, we can divide up hypothetical sound patterns into two groups: those that seem like natural, garden-variety generalizations, and those that are highly unnatural and involve patterns that are really unlikely to be found across languages. Remember that in Chapter 2 we saw that the existence of language universals and tendencies has been used to buttress the argument that much of language learning is innately constrained. The idea is that children come pre-equipped with biases to learn certain kinds of linguistic patterns over others, and that this is why we see evidence that some patterns are more common across languages than others.

Do babies start life with a bias for certain kinds of statistical regularities in the sounds of speech? If they did, they could avoid wasting their attention looking for patterns that just don’t seem to be useful for languages. We might predict, then, that they’d be very unlikely to notice the generalization about word-final /m/ and /s/ being dependent on word-initial /k/ or /l/. A statistical rule like this fails a “naturalness” test on four counts: First, the sounds to which the rule applies—/m/ and /s/—don’t form a natural class of any kind, making them odd bedfellows for a rule. Second, the sounds that characterize the relevant linguistic context—/k/ and /l/—are even odder companions, not even belonging to the same category of sound at the broadest level (one is a consonant and the other is a vowel). Third, there’s a yawning distance between the word-final sounds and the linguistic contexts on which they depend. And fourth, the relationship between the word-final sounds and the word-initial sounds that determine their identity is purely arbitrary. All of this hardly makes for a promising statistical pattern.

To find out whether babies strategically allocate more of their cognitive resources to statistical hypotheses that are most likely to pan out for a natural language, we’d need to test a large number of natural and unnatural patterns embedded into artificial languages. But there’s at least some evidence now that shows that not all statistical patterns are learned with equal ease. Jenny Saffran and Erik Thiessen (2003) compared two kinds of phonotactic constraints: In one experiment, 9-month-old babies had to learn that the first sound in a syllable was a voiceless stop (/p/, /t/, /k/), while the last sound was always a voiced stop (/b/, /d/, /g/). That is, the phonotactic constraint applied to a natural class of sounds. The babies showed signs of learning this pattern after hearing a speech sample of only 30 words, repeated twice. In another experiment, babies had to learn that syllables began with /p/, /d/, and /k/ and that they ended with /b/, /t/, and /g/. In this case, voiced and voiceless stops were mixed together as possible sounds at both the beginnings and ends of syllables, and the babies would have to learn the statistical rules in terms of individual sounds, and not in terms of natural classes of sounds. The end result was that, after the same amount of exposure as in the first experiment, there was no sign that the babies were picking up on this pattern.

So, there are some intriguing results showing that highly natural patterns leap out at babies more readily than unnatural ones. One way to think about this is that the learning biases shown by these tiny language learners correspond to innate “settings” that constrain them from
generating wildly unhelpful hypotheses about the structure of language—in other words, they correspond to a type of innate knowledge about natural language patterns. But there’s another very different explanation that fits with the same experimental results. As you saw in Chapter 2, the nativist argument based on language universals can be flipped on its head. It may be that, instead of reflecting an innate program that guides the process of language acquisition, language universals reflect a process of languages adapting to the limitations of the human mind. In other words, it may be the case that some kinds of statistical regularities are simply easier to learn than others, or need less mental horsepower to compute. Patterns that are hard for children to learn eventually get weeded out of a language. This explanation is quite a plausible one, especially given that we’ve already seen that languages tend to adapt to basic properties of the auditory system—for example, like voicing in English, they might shape their phonemic inventories around sound distinctions that are especially easy to perceive.

Still, before we can really pull apart these competing ways of looking at language universals and learning biases, more work needs to be done. We need to establish what kinds of patterns are more easily learned than others, and why. This means building up a body of knowledge about “easy” versus “hard” kinds of patterns, and looking at how these play out across species, across various domains, and perhaps across the developmental span of humans. (Remember, as I hinted at in Chapter 2, the ultimate shape of a language may be influenced by the age at which most of its learners typically acquire it.) The more similarities we see across domains and species in terms of easy versus hard patterns, and the more that the distinction between easy and hard patterns aligns with language universals, the more evidence we have that statistical learning biases are deeply embedded within more general cognitive skills rather than reflecting an innate, language-specific program. On the other hand, if language-related biases turn out to be dramatically different from the kinds of biases we see in other species and other domains, this would provide some support for the idea that there are specific and possibly innate constraints on learning the sound patterns of human language. Care to lay any bets on how it will turn out?

**PROJECT**

Think of a statistical pattern that seems unlikely to be a natural pattern in a language. Create a snippet of an artificial language, and formulate a research design in which you would test to see whether subjects are sensitive to this particular kind of statistical information. If possible, test a group of adult subjects, and analyze the data. For a description of the artificial grammar studies with adult subjects, see Saffran et al. (1997).