Chapter 6

Words

chapter 6

THE FIRST STEP IN understanding both spoken and written language is doing something with words. Words are the building blocks of language.

I don't want to claim that we necessarily understand language on a strictly word-by-word basis. Sometimes we might use information from other words to identify some words. If you hear "The cat chased the [slurred speech sound] -ouse", you'll probably decide the final word is "mouse" rather than "house" or "louse". The extent to which surrounding information influences the identification of words is one of the most important controversies in the field. But to get to the point of wondering whether it's a mouse or a louse, you have to have identified most of the other words in the utterance. We say that word recognition is primarily a bottom-up or *data-driven* process. We're trying to understand what's in front of us, visually or aurally, primarily on the basis of what's in front of us. The alternative to data-driven processing is top-down processing, where we use general knowledge or hunches about what we're processing to identify it. Although there might be some role for it, clearly top-down processing by itself would be a terribly inefficient way to understand language. We're not going to get far by ignoring what the other person says and claim we know what they mean because we

think we do (although in my experience many people act this way). So recognising words and understanding language has got to be mainly data-driven. The question is what is the role of top-down processing – if any.

For once we don't need a dictionary to tell us what a word is, but the rest of the process isn't as clear. Note that I began by saying "doing something with words". I've also been a bit sneaky using words such as "recognise", and "understand" and "identify", without really explaining them. It's time to unpack these ideas a little. Let's think about our goal when listening or reading: it's getting enough meaning from the words to be able to construct a representation of the meaning of the sentence, which we can then use to do something with. *Recognising* a word means you've made a decision in some way that the word is familiar; you know that NIGHT is a word and you've seen it before, and you know that NITE isn't. Strictly speaking, recognition doesn't necessarily entail anything more: you could in principle decide something's a word you know and not do anything more with it. *Identifying* a word means that you've made some commitment to what the word is – sufficiently so to be able to initiate some response. *Understanding* a word means that you access the word's meaning. *Naming* a word is accessing the sound of a word, which in turn could mean saying it aloud, or saying it to yourself. When we're reading, do we automatically and necessarily access the sounds of the word? And then there's a term much liked by psycholinguists, *lexical access*: that means accessing our mental dictionary, the lexicon, and obtaining potentially all knowledge about the word – its meaning, sound, appearance, and syntactic information about it.

Psycholinguists have used several tasks to investigate word processing, and these tasks relate to these distinctions I've just made. One of the most popular tasks is the *lexical decision task*; imagine you're sitting in front of a computer screen, and a string of letters is flashed up in front of you. You have to press one key if you think the string of letters forms a word, and another if you think it's a non-word. So you might see NIGHT or NITE. Researchers measure how long it takes the person to make their decision, and also how many errors they make. In a lexical decision, you don't have to access the meaning or the sound of the word, you just have to say whether it's familiar or not – whether it's in that set of things you know to be words; you *might* access the sound or meaning, but you don't have to. Contrast lexical decision with the *naming task*, where you see a word on the screen and have to say it aloud, and researchers measure how long it takes you to start speaking. In naming all you have to do is pronounce it, which means that you don't *have* to access the meaning or even decide whether it's familiar or not; and even if you do access the meaning, we don't know that it happens before you start naming. There are other tasks, but these two illustrate the difficulties involved in talking about word processing: access to meaning, the sound of a word, and its familiarity could happen at different times, and we have to be clear about what we think we're measuring.

Many researchers believe that there's a "magic moment" in word processing, where a person has recognised a word but hasn't yet accessed the meaning. Put slightly more formally, this is the point at which a word becomes sufficiently activated for a person to carry out some sort of response to it, but this point is before we start to access the meaning. As Balota (1990) points out, this assumption sounds like a reasonable one, and at first sight how could it be otherwise? Surely we have to identify the word before we can access its meaning? No, we don't have to, because we could start accessing something about meaning as soon as there's some information about the word coming through, and this meaning might be used to influence the identification process. If this idea is correct, then we have to be very careful about what we conclude from lexical decision and naming, because these judgements might be influenced by meaning.

You can tell by now that word processing isn't going to be straightforward. We have to be very careful about what tasks we use to study it and what assumptions we make. We can't assume that when we start naming a word we've accessed its meaning, or that we can decide whether something is familiar or not before we access the meaning. Even "accessing the meaning" is a phrase full of pitfalls. When we hear or read a word, like TIGER, how much of its meaning do we access? Must we access everything we know about it? Do we automatically and necessarily retrieve how many legs it's got and that it's got stripes and whiskers? I think two features of word processing make our life easier: most of the time it's just "good enough", so we only do as much with the incoming words as needed to get by on, and it's got a statistical or probabilistic element to it so that, some of the time at least, we're almost just guessing. One problem with the area of research into word processing is that it's been dominated by metaphors that we have come to believe to be true. We talk about searching the lexicon as though we're looking through a dictionary, and then we have models of lexical access based on a dictionary search.

We know that many factors can influence the ease with which we can recognise a word (Harley, 2008). Obviously the clarity of the perceptual stimulus matters; it's going to be more difficult to recognise a very quiet, mumbled word spoken against a lot of background noise than a nicely enunciated loud one spoken against a background of silence. As I've mentioned before, the frequency of a word is an extremely important variable in word processing – the more common a word is, the easier it is to recognise. It's thought that the age at which we are first exposed to a word – a variable called age-of-acquisition – is independently important, such that we're faster to process words we learn earlier. Having seen a word in the recent past makes it easier to identify. Finally, I must mention semantic priming – we find it easier to identify a word if it is preceded by one related in meaning (such as DOCTOR and NURSE).

How do we recognise spoken words?

Have you ever tried listening to someone speaking a language you don't know? You can't even make the words out. I find it difficult even with a language I know a tiny bit about, like Spanish or French; listening to a native speaker talking at their normal speed is a chastening experience. I can occasionally make out the odd word, but mostly it's just a string of sounds. It's the same problem babies face when learning language: segmentation. The sounds of speech are usually slurred together. Of course there are some pauses, and small gaps after some sounds (those called stop consonants, that necessitate closing the airstream very briefly, such as p, b, t, d), but although speech is rapid, we're very effective at processing it. If we're given a mixed-up sequence of buzzes, hisses, tones, and vowels, we can only distinguish the order of the sounds if they come at a rate slower than 1.5 sounds a second; but we can understand speech at the rate of 20 sounds a second, and sometimes faster (Warren & Warren, 1970). We're good at processing speech.

The basic unit of speech is called the *phoneme*. A phoneme is the smallest unit of a language that makes a difference to the meaning. So the sounds "p", "b", "r", and "c" are all phonemes, because if we swap them around we get different words with "pat", "bat", "rat", and "cat". By convention, we put phonemes in sloping lines, like this: /p/. Phonemes vary from language to language; /l/ and /r/ are different phonemes in English (as testified by the difference between "lot" and "rot"), but not in Japanese. Other languages make distinctions that English doesn't. You might want to try the following in private. Put a hand in front of your mouth and say the word "pin"; you should be able to feel a puff of air accompanying the "p" sound. Now say "spin"; there isn't any such puff. Physically they're different sounds – we say the "p" in "pin" is aspirated and that in "spin" is unaspirated – but the difference isn't a critical one in English. In some languages, such as Thai and Korean, it is; whether or not the "p" is aspirated can change the meaning of the word.

This point illustrates the fact that a phoneme can vary considerably, which is another factor that makes speech processing difficult. To make matters even more complex, the precise sound changes depending on what other sounds it's surrounded by: the /b/ phonemes in "bill", "ball", "able", and "rub" are all acoustically slightly different. This phenomenon is called *co-articulation*. It happens because as we say any sound, the articulatory apparatus is moving, getting ready to produce the next sound (and indeed has moved into position having just produced the previous one). Co-articulation could be of some assistance, because it means that a sound gives information not just about itself but also about surrounding sounds, but it is another source of variability in sound. And of course, no two people speak in exactly the same way – in addition to systematic differences such as age, gender, and regional accent, there are individual differences. All these ways in which sounds vary depending on the context also make speech recognition very difficult. They rule out models of sound recognition based on templates, where we

compare each sound to an internal idealised phoneme, because the incoming sounds are just too variable.

Even though speech is highly variable, we're not that sensitive to the differences. Of course we can identify the gross characteristics of the speech, enough to be able to identify them, their age, and gender, but we don't hear all these variations in sound as different phonemes. To the English ear an aspirated and unaspirated "p" are just /p/s. And what happens if we hear a sound that's intermediate between two sounds? We categorise it as one thing or another. We simplify what we hear.

It's worth dwelling on this point about intermediate sounds a little more. Let's take as an example pairs of sounds like /p/ and /b/, and /t/ and /d/. The words in these pairs lie on the ends of a continuum. The difference between the ends of the continuum, between a /p/ and a /b/, is called their *voice onset time*. With *voiced* consonants (/b/, /d/), the vocal chords start vibrating (you can feel them vibrate if you put your fingertips to your throat) as soon as the lips close (/b/) or the tongue goes to that little ridge above the teeth (/d/); the voice onset time for voiced consonants is close to 0 milliseconds. With voiceless consonants (/p/, /t/) there's a bit of a delay (60 milliseconds or so) before the vocal chords start vibrating. This small difference in voice onset time is all that separates a /p/ from a /b/ and a /t/ from a /d/. But what happens if you hear a sound halfway between, with a voice onset time of 30 milliseconds? Do we hear a sound halfway between a /p/ and a /b/?

No, we don't, we hear either a /p/ or a /b/; there's no halfway house, in that all variants of the same phoneme sound the same to us, a phenomenon known as *categorical perception* (Liberman et al., 1957). Which we hear varies from person to person and depending on the circumstances. The closer the voice onset time to the 0 milliseconds end of the continuum, the more likely we are to categorise it as /b/, and the nearer to the 60 milliseconds end the more likely we are to categorise it as /p/, but there's no in-between stage where we hear something halfway. If we go up the onset time continuum we switch suddenly from hearing /b/ to /p/. And the same is true of /d/ and /t/. Although there is a distinct boundary between the categories, it isn't fixed. We can move the boundary up or down by fatiguing the feature detectors that identify the sound by repeated exposure to a sound from one end of the continuum; so if you hear /p/ repeated several times, you become a bit more likely to identify a sound halfway along the continuum as a /b/ (Eimas & Corbit, 1973). The categorical perception of sounds is probably a result of the way in which the brain is wired rather than being a skill we learn; babies as young as one month show it (Eimas et al., 1987). It isn't even unique to humans: chinchillas, a cute sort of little South American rodent, categorise syllables such as "ta" and "da" in just the way humans do (Kuhl & Miller, 1975). The phenomenon of categorical perception is not even restricted to speech, as musicians appear to perceive musical intervals categorically (Siegel & Siegel, 1977).

Another possible source of assistance when listening to speech is the knowledge we have about what the speaker might mean – what we call the context. The

importance of context is demonstrated powerfully by a famous psycholinguistic phenomenon known as the *phoneme restoration effect* (Warren, 1970; Warren & Warren, 1970). You hear the following sentences:

It was found that the *eel was on the orange.
It was found that the *eel was on the shoe.
It was found that the *eel was on the axle.
It was found that the *eel was on the table.

They're constructed by splicing tapes together (so easily done digitally these days; I wonder at the perseverance of researchers having to do these experiments without the full panoply of modern computing tools) so that the only word that differs is the final one. The asterisk represents a 0.12 second portion of speech that's excised from the tape and replaced with a cough. People then say they hear "peel" in the first instance, "heel" in the second, "wheel" in the third, and "meal" in the fourth. They don't hear there's anything missing, and can't even reliably locate the cough in the right place. People say they actually hear the sound that isn't there. The only explanation for this is that the "top down" knowledge of the word is affecting the perception of the speech stream. The precise sound doesn't matter – buzzes and tones elicit the effect just as well, and you can excise larger amounts of the word and still get the effect. There are limits on how much and when people will restore, though: you can't get away with just replacing the phoneme with a short period of silence; people notice that.

The influence of context on speech perception isn't restricted to speech; what we see can influence what we hear, as demonstrated by the McGurk effect (McGurk & MacDonald, 1976). Suppose the sound "ba" is played to you through headphones while you simultaneously see a video of someone saying the sound "ga". What you hear isn't the sound "ba" at all, but a sound in between the one you should hear and the one you hear": "da". The McGurk effect demonstrates an interaction between visual and auditory perception; we are using information about the shape of a speaker's lips to deduce (of course, not consciously) what sound they are producing, and this information in turn influences what we hear.

So although there's a great deal of variability in speech, we have ways of simplifying that variability. We just deal with categories of sound, rather than every subtle shade of variation, and use the context to narrow down the search.

I want to examine briefly a couple of models of speech recognition. They share the underlying idea that when we hear speech, the sounds of words activate words that could possibly correspond to the perceptual input. I said earlier that the concept of activation is a powerful one, and it pays its way here.

The *cohort model* of William Marslen-Wilson emphasises the way speech unfolds across time (Gaskell & Marslen-Wilson, 2002; Marslen-Wilson, 1990; Marslen-Wilson & Welsh, 1978). You're listening to a word, and of course you

don't hear it all at once, you hear it sound by sound. So suppose you're sitting in a lovely drawing room watching an old lady sewing, and she says the following:

> Be a dear and pass me some thread for my –

Now what could come next? The syntax provides some constraints – only an adjective or noun makes sense. Context provides others; it's most unlikely that the word coming up is going to be "elephant". So already the *cohort* of candidate words that could be next has been greatly reduced. Next you hear the first phoneme, /t/. This small amount of perceptual information makes an enormous difference and can eliminate tens of thousands of words from the cohort of candidates. What could the word be? I can only think of a few plausible candidates that could now be left in the cohort, "tablecloth", "tapestry", "togs", "tea cosy" perhaps, although some of these are less likely than others. If the next sound is "a" (as in "tapestry"), I think we're really only left with "tapestry" remaining in the cohort – the context precludes the possibility that she might be asking for some thread for her tapeworm. In this way the elimination of candidates from the cohort can be rapid, and therefore spoken word identification is very fast and effective.

All spoken words have a point at which they become unique – that is, based on perceptual information alone, we can say with absolute certainty "this word I'm hearing is tapestry, not tapeworm or tabulation". That point is called the *uniqueness point*. We can find out what each word's uniqueness point is using what's known as the *gating task*, in which participants are played increasingly large segments of a word, perhaps in 20 millisecond slices, and asked to say what word they're hearing. They start getting it right after the uniqueness point. The gating task shows the importance of context; people need on average 333 milliseconds to identify a word in isolation, but only 199 milliseconds to hear a word in an appropriate context. So we're slow identifying "camel" in isolation, but much faster when we hear it in "At the zoo, the kids rode on the camel" (Grosjean, 1980).

What's in the cohort matters. If what you're hearing has a very unusual phonological form ("xy-"), the cohort of words is going to be very small, but if you hear something that has a common phonological form ("sp-"), the cohort of candidates is going to be very large. It's hardly surprising therefore that cohort size affects word recognition time; if the cohort is large, recognition is slower, and what's more, the relative frequency of all these neighbours matters. If the target word is "specious", it's as though the common words like "special" and "speech" are getting in the way. So you're faster to recognise a high-frequency word that has only low-frequency neighbours than vice versa (Goldinger et al., 1989; Marslen-Wilson, 1990).

This broad approach of activation-based models of word recognition has been implemented in a computer simulation known as TRACE (McClelland & Elman, 1986). In the TRACE model there are three levels of processing units. Each unit is very simple, simply accumulating activation from all the other units to which

its connected and passing some of it on to other units. At the lowest level is a pool of phonetic features, in the middle level is a pool of phonemes, and at the highest level is a pool of words. Phonetic features are components of phonemes; for example, as we've seen, /b/ has a voiced feature, /p/ a voiceless feature; consonants are produced by constricting the vocal tract at some point; some sounds are strident, containing a burst of white noise, such as /s/ and /sh/. The details of these phonetic features needn't detain us; the important point is that it's possible to decompose phonemes into lower-level units, and specify how two phonemes differ in terms of their featural make-up. /b/ and /p/ differ, for example, by the presence and absence of voice.

In line with the principle that connectionist networks display massive interconnection, every unit in one level is connected to every unit in the level above. However, there are three important points about these connections. First, they're bidirectional, which means that activation can flow down the network from the word level to phonemes and phonetic features as well as bottom up. Second, they're either excitatory or inhibitory. An excitatory connection is one that increases the activation level of the unit at the other end, and an inhibitory connection is one that decreases the activation level at the other end. So the phoneme /t/ has excitatory connections to TAKE and TASK, but inhibitory connections to CAKE and CASK (Figure 6.1). Third, each unit within a level is connected to all other units in that level by an inhibitory connection. All we have to do to get the model rolling is to apply activation to the input phonetic features corresponding to a word, and sit

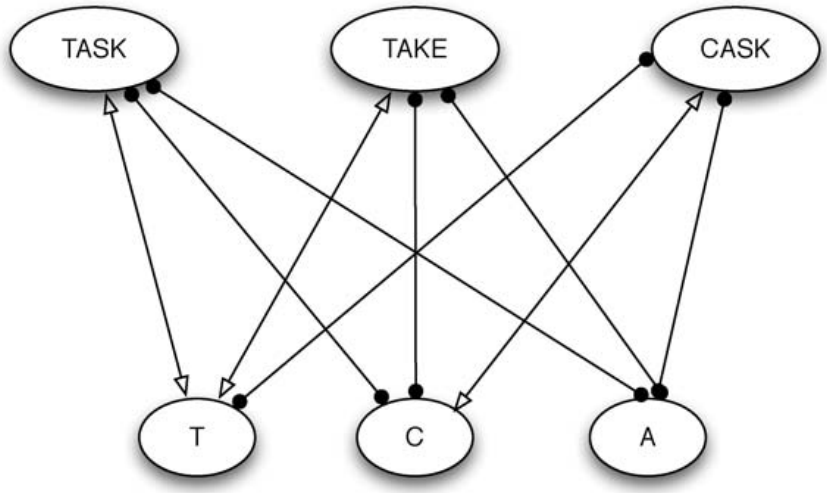

FIGURE 6.1 Portion of a TRACE-style network. This fragment simply shows some of the connections for the initial-letter position. T is in the first position of both TASK and TAKE, so the T-initial unit is connected to these units with bidirectional facilitatory connections. It isn't in CASK, so T is connected to that unit by an inhibitory connection. The reverse applies for the C-initial unit. A isn't in the initial position in any of these words (although it is of course in the second position), so the A-initial unit is connected with inhibitory connections

back and watch the activation flow up and down the network until only one word unit is left active – and that's the word the system has "recognised". Suppose we apply activation corresponding to the phonetic features unvoiced, consonant, tongue touches above the teeth; these are enough to get the /t/ phoneme going. This activation then spreads to TAKE and TASK, and these will inhibit incompatible patterns, such as CAKE and CASK. When an /s/ sound comes along a bit later, TASK will start inhibiting its competitor TAKE. Eventually only one word is left standing. Models like TRACE are called IAC models, short for interactive activation and competition. Competition is an important aspect of this network; as soon as a unit starts to pile up evidence for itself, it starts to inhibit its competitors. Truly do the rich get richer; the winner takes all.

Because activation is top-down as well as bottom-up, the model can deal with incomplete or corrupted evidence. Suppose the model is presented with a phoneme intermediate between /p/ and /b/ – the voicing feature is given 50% activation – but that is followed by -LUG. Only PLUG is a word, so that gets activated, and sends activation down to the /p/ unit, which then sends activation back down to the featural level. BLUG is not a word so /b/ doesn't get much of a look in, even though /p/ and /b/ sound very similar (differing in just the voice feature). So context exerts its influence by top-down activation spreading to lower levels. And because there's always got to be a winner, the model abhors intermediate values. It's got to be either a /p/ or a /b/ we're hearing, and one of them has got to win, so the model displays categorical perception of phonemes as well.

These models demonstrate the power of activation-based models and the usefulness of implementing your model in a computer program to show that the model really works. Often things drop out of the simulations you don't expect. Phenomena such as phoneme restoration and categorical perception of phonemes arise from the principles of top-down activation and competition; the explanations make sense, but there's nothing like demonstrating it in a computer program for reassurance. Needless to say, in psycholinguistics no model goes unchallenged. The main bone of contention here is exactly when context has its effect. Does it directly affect the very earliest, perceptual stage of processing, or much later, when we're trying to integrate what we've perceived with everything else? Is phoneme restoration a true perceptual effect, or does it reflect guessing? Can we trust what people report? The East Coast view is that perception is bottom-up, with context affecting only the later stages of processing, while the West Coast view is that context can affect early processing. The jury's still out.

The TRACE model as implemented is just a fragment of presumably what actually happens. The word units will in turn be connected in a similar way to the semantic feature units; meaning will seep back down to the word influence to supply activation to contextually appropriate words, which in turn will flow back down to the letter level, and so on. In this kind of model where activation is continually "cascading" through the system – where as soon as one unit sees evidence for itself

it immediately sends out activation to all the other units to which it's connected – there can be no magic moment where we recognise a word but don't access any of its properties.

I've also been a bit vague about what "context" is. Potentially it's all the knowledge we have that could influence perception top-down. So if we're trying to identify a phoneme, it's the context of what word it could be, how the possible words fit in with the syntax of what we're hearing, our knowledge about the conversation, the situation, the weather – anything at all that could help narrow what's being said or read. And that's the problem with context: it's huge. And do we really want to override our perceptions too easily? If a lion bursts through my office door, should I act on the basis of my perception and run, or should I stop and reason with myself about the unlikeliness of this event and therefore how my perception could be wrong, while I get eaten?

How do we read?

I probably spend more time reading than talking, but I realise I'm almost certainly unusual in this respect. Writing has had a huge influence on culture and thought; by

Not the best time for a lot of top-down processing

Don't try this at home: cascading activation. Water is already pouring into the bowl even while it's still flowing out of the top glass

its enabling us to create and store external records of our thoughts and memories, we have increased our cognitive capacity enormously. It's impossible to imagine that civilisation could have progressed far, and that we could have developed any sophisticated technology, without written language. Indeed, the invention of writing is what separates history from prehistory.

Nevertheless, students often have the impression that a disproportionate amount of psycholinguistics is taken up with two apparently very specific topics: how we understand the sort of sentence known as the "garden path", which I'll come to in the next chapter, and how adults read single words. I used to worry about this restriction too, before, with respect to Dr Strangelove, I learned to stop worrying and love single-word recognition. If we understand how we read one word we know something about how we read them all. There is also the advantage that it's relatively easy research to carry out; these days all you need is a computer and you can carry out sophisticated priming and lexical decision experiments. And because it's easy to carry out, there's been a lot of research, so we know what's

A more conservative solution: discrete processing. The tall glass is empty; the short one is just about full and ready to pour in one go into the empty bowl

happening, and if we don't know the answer, we know the shape of the battleground, so word recognition exemplifies some of the key topics in psycholinguistics. Of course reading does involve much more than reading isolated words, so it's right not to get too carried away. In real life, we have an array of text in front of us, which means that we can jump around the text, looking back to earlier material if need be. It also means that the eye can take in information about more than one word at a time, which could be of considerable assistance in understanding them. There's now a considerable amount of evidence that when reading we can take in information about words that fall on the retina outside its most sensitive spot, the *fovea* (e.g. Kennedy & Pynte, 2005).

Recognising printed words presents a different set of problems from understanding spoken words. The words are usually (except in devious psycholinguistic experiments) fixed and unchanging in front of us for as long as we need. But we have to decide where to put our eyes to extract information from the printed page in a fairly efficient way. One thing we do know for certain is that in normal adults (those without brain damage), reading is like listening; you can't help yourself – you have to do it. You can't choose to stop the reading process halfway through once you've seen a word, in the same way that when you hear a word you can't help but

understand it. That reading is mandatory is shown very clearly by the Stroop task (Stroop, 1935), in which you have to name the colour of ink a word is printed in. That sounds easy enough, but if the word is a colour name, and is different from the ink colour, you're much slower than when the name and ink colour are congruent. So naming the colour in RED (in red ink) is easy, but GREEN (in red) is hard. You can't stop yourself reading the word, and you can't stop the meaning interfering with what you're supposed to be doing.

Unlike speech, which could reasonably only have evolved once, writing seems to have been invented several times in man's history, almost certainly arising in the near and middle east, China, and the Mayan culture of ancient Mesoamerica. As speech might have evolved from gesture, writing almost certainly evolved from some simpler system, probably a means of keeping tallies of numbers.

I II III IIII IIIII

Those languages in the near and middle east originally used pictures to represent concepts, with the first probably being the cuneiform script of ancient Sumer around or before 3000BC. The well-known hieroglyphic system of ancient Egypt probably developed soon after, with the hieroglyphic system of Crete developing after that. These early systems used pictures to convey meaning, although the relation between the picture and the meaning gradually became looser, with some symbols coming to stand for sounds.

The big invention that changed western writing was the widespread adoption of the alphabetic principle, where the written symbols stand for individual sounds. The alphabet probably originally derived from the Egyptian system of hieroglyphs, but it became a central part of the Phoenician system of writing. As Phoenicia, located approximately in the modern coastal territories of Syria and Lebanon, was an important maritime trading state, the alphabet spread throughout the eastern Mediterranean. The Phoenician alphabet represented just the consonants of the language; the Greeks adopted it and added vowels. The Roman alphabet subsequently took over the Greek system, and the Roman Empire spread the alphabetic system throughout the western world. Today languages in the west are based on the Graeco-Roman alphabet, with those of China, Korea, and Japan based on the ancient Chinese system of pictures.

A legacy of this complex and multi-centred evolution is that today different languages use different means to map written words on to language. In a language such as Chinese, every word has a different symbol or combination of symbols associated with it, and you (more or less) have to learn each word separately. Such languages are called *logographic* languages. There are over 45,000 symbols in the full Chinese dictionary, although most of these are rarely used and so full literacy is possible with knowledge of under 4000. Contrast this complexity with the languages that use the *alphabetic* system, where each letter corresponds to a sound:

English of course uses just 26 letters. Even then there are different ways of mapping letters onto sound; the consonantal scripts of Hebrew and Arabic continue to represent just the consonants, with the vowels being filled in when reading. Most alphabetic languages – including English – represent both consonants and vowels. But even within these alphabetic languages, the details of the ways in which letters correspond to sounds differ. In some languages, such as Serbo-Croat, each letter corresponds to just one sound, and vice versa. We call these *regular* languages – or sometimes ones with shallow orthographies. In languages such as French, correspondences between letters and sounds are regular, but some sounds can be represented by different combinations of letters ("o", "eau", "au", "eaux", for example). In English the relation between sound and letters is complex, with some sounds being represented by different combinations of letters (consider for example "to", "too", "two", and "threw"), and some letters corresponding to different sounds (consider the "a" in "hat", "hate", and "father"). This lack of regularity makes English spelling particularly difficult to learn.

What's the dual-route model of reading?

"Xhjhhgz" is a non-word, but it's not a very interesting one. You can't pronounce it. Instead, let's take the examples of "gat", "smeat", and "nouse". You can pronounce – say aloud – these; they could be English words, but they happen not to be. We call pronounceable non-words like "gat", "smeat", and "nouse" *pseudowords* – they're like words, but aren't quite. The comedian Stanley Unwin made a decent living speaking in these, using his language Unwinese to comic effect (Elvis Presley was rather beautifully described as "wasp waist and swivel hippy"). What's more, if you ask a number of people to pronounce these pseudowords, they usually all give the same pronunciation. The most obvious explanation of how people can read things they've never seen before is that they build up the pronunciations, letter by letter, converting each letter into a sound, and assembling those sounds together.

Now consider a word like BEEF. Imagine you've never seen it before. How would you pronounce it? You could do just the same as with these pseudowords: you could convert letters into sounds and say B EE F – and your pronunciation would be quite right. Words like BEEF, where all the letters are given their most common pronunciations, are called *regular*.

Let's turn BEEF into STEAK. How would you pronounce that if you'd never seen it before? The common pronunciation of EA is "ee", as in "speak", "leak", and "bleat". So starting from scratch you'd get STEAK wrong, and you'd pronounce it "steek". Words like STEAK, where the letters or letter pairs don't have their most common pronunciations, are called *irregular* (or exception) words. English is full of irregular words. "Steak" is by no means the worst. What about "aisle", "ghost", "psychology"? You might have heard the following old joke:

How do you pronounce the word "ghoti"?

Pause for a moment if you haven't seen it before. The answer is "fish" (with "gh" as in "enough", "o" as in "women", "ti" as in "rational".

Here then are two critical observations about reading English. We can pronounce irregular words and we can pronounce novel pseudowords. The obvious explanation for these abilities is that reading involves two processes: we read pseudowords (and, in principle, regular words) by a process taking the word letter by letter, and turning each letter into a sound, a process we call *grapheme–phoneme conversion* (GPC). Grapheme–phoneme conversion won't work for irregular words. We just have to know them all by heart.

So without too much difficulty we've created a model of reading aloud, the dual-route model, championed across the years particularly by Max Coltheart (Coltheart, 1985; Coltheart et al., 2001). The dual-route model posits two routes from print to sound. First, there is a fast, direct route, where print activates an entry in our mental dictionary, the lexicon. When the entry is activated, we gain access to all information associated with the dictionary entry, including the word's meaning and sound. The direct route, called the *lexical route*, is fast and effective for skilled readers and well-learned words. Second, there's an indirect route, making use of letter–sound correspondences, called the *non-lexical route*. For skilled readers this route is much slower, but novice readers depend on it, using GPC to spell out the sound of a word and then using that sound to access the lexicon. We can think of the two routes in a perpetual race, with the lexical route becoming faster the more skilled a reader we become. We're very fast at reading regular words because there's no conflict between the direct and indirect routes, but with irregular words the two routes give conflicting answers, and that conflict slows us down, explaining why we're faster at reading regular words than frequency-matched irregular words (Baron & Strawson, 1976).

What does brain damage tell us about reading?

If there are two reading routes, it seems very plausible that these might be located in different parts of the brain. It follows that, by chance, some people will have damage to the part of the brain housing one route but not the other, and with other individuals the opposite should be the case. Put more concretely, we should find some adults who as a consequence of brain damage should have damage to the lexical route but not the non-lexical route, and other adults who have damage to the non-lexical route but not the lexical. Such a pattern, where two skills can be differentially impaired, is called a *double dissociation*. Note that I'm just talking about adult *acquired* dyslexia here – adults who could previously read well but suffer brain damage, often as a consequence of a stroke, that affects their reading

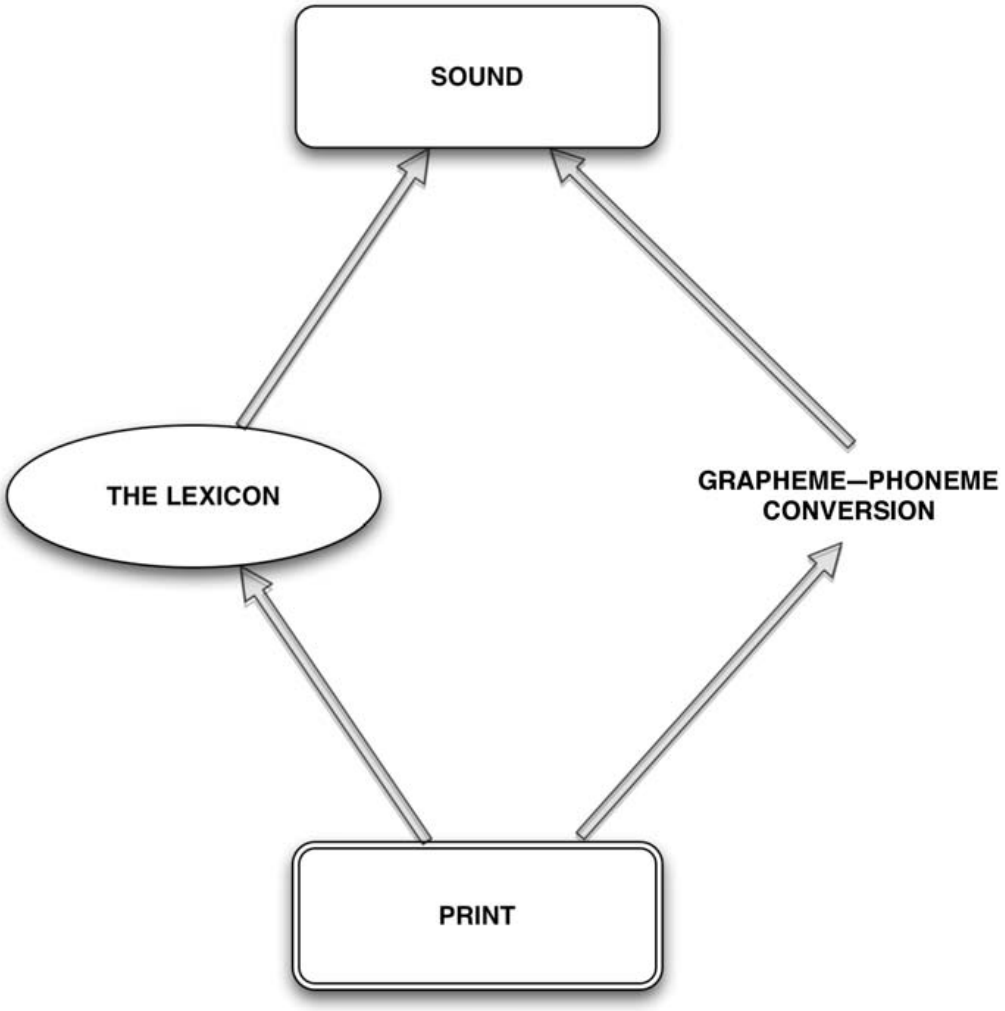

FIGURE 6.2 The dual-route model of reading

ability – in contrast to people, particularly children, who have difficulty in learning to read, a difficulty called *developmental dyslexia*.

What would such patterns look like? Remember that the lexical route is a direct route, where we access the pronunciation of a word directly, and is essential for the correct pronunciation of irregular words. People with damage to the lexical route will therefore have difficulty reading irregular words, but should be able to manage with regular words and pseudowords, because they can be read by the sublexical route. This pattern of impairment exists and is called *surface dyslexia* (Marshall & Newcombe, 1973). Patient MP (Bub et al., 1985) was almost perfect at reading non-words, showing that her sublexical route was completely preserved, but she was poor at reading irregular words, managing to read correctly 85% of high-frequency irregular words and just 40% of low-frequency irregular words, suggesting that the sublexical route was severely damaged. Surface dyslexics make

just the type of errors you would expect if you try to read an irregular word through the sublexical route using grapheme–phoneme conversion; they try to regularise the irregular words, trying to pronounce them as though they were regular, hence producing errors such as "steek" for "steak", "eyesland" for "island", and "brode" for "broad". These are called *regularisation errors*. Put at its most dramatic, a person with surface dyslexia could read the regular word "speak", but not the very similar but irregular word "steak".

What about damage to the non-lexical route when the lexical route is left intact? People with this sort of impairment should be able to read words (e.g. SLEEP) but would be unable to read pseudowords (e.g. SLEEB). This pattern of reading is known as *phonological dyslexia* (Shallice & Warrington, 1975). A patient known as WB is a very striking case of someone with phonological dyslexia (Funnell, 1983); WB could not read any non-words at all, suggesting that the GPC sublexical route was completely obliterated, but could read 85% of words, indicating that his brain damage had almost totally spared his lexical route.

There are a few puzzling observations about phonological dyslexia, however, that make its interpretation a little more complex. First, for that vast majority of patients who have some sparing of the sublexical route, their reading performance is generally much better if the non-word when pronounced sounds like a word. Such pseudowords, called pseudohomophones, include as examples NITE, BRANE, and FOCKS. The most obvious explanation for this finding is that there is some sort of leakage between the lexical and non-lexical routes. Another finding is that phonological dyslexics tend to have particular trouble reading low-imageability low-frequency words, so it should come as no surprise that they find grammatical words particularly hard to read. They also have difficulty with word endings, finding inflected words difficult to read. Phonological dyslexia resembles deep dyslexia, but without the semantic paralexias; indeed, deep dyslexia sometimes resolves into phonological dyslexia as the patient recovers some ability after the brain trauma. Put at its most dramatic, a person with phonological dyslexia could read the word "sleep", but not the very similar non-word "sleeb".

Given the existence of these two types of dyslexia resulting from brain damage, it follows that different regions of the brain must be involved in different aspects of reading. Further evidence for specialisation of brain regions for different aspects of reading comes from a brain imaging study by Fiebach et. al (2002). They examined the way the brain becomes activated when it has to make lexical decisions as to words and pseudowords. Different parts of the brain lit up more depending on which type of stimulus was presented: words elicited more activation than pseudowords in the part of the brain around the bump known as the temporal gyrus, whereas pseudowords elicited more activation in more frontal and subcortical regions. Do the imaging data necessarily support the dual-route model? While the data are consistent with it, it would be too strong to say that they confirm the dual-route model. They're consistent with any model of reading that has at least two

processing components. As we'll see, the triangle model says that there's a division of labour between the orthographic–phonological and semantic pathways. So as it stands, the imaging and neuropsychological evidence can't distinguish between the models.

What are the problems with the dual-route model?

So far, so good; the data support the dual-route model, although some of the observations concerning phonological dyslexia are a little unsettling. But it soon became apparent that we can't get away with a model of reading that's that simple. The three main obstacles to a simple life are that there are lexical effects on non-word reading, not all sorts of word are equivalent, and there are other types of acquired dyslexia that don't fit so well into the basic model.

That not all non-words are created equal was demonstrated in an ingenious experiment by Glushko (1979). Glushko examined the effects of word neighbours on pronouncing non-words. A neighbour is a word that's very similar to a word or non-word; for practical purposes we can say that two words are neighbours if you turn one into the other by changing one letter, so by this definition "gaze", "maze", and "laze" are all neighbours. Now consider the non-word TAZE, which has all those words as neighbours. What's important about these neighbours is that they're all regular words. Contrast that with the non-word TAVE; that has lots of neighbours, including the regular "gave", "save", and "rave", but it also has an exception neighbour, the frequent and potent "have". People are faster to name non-words with consistent lexical neighbours (like TAZE) than non-words with inconsistent, squabbling neighbours (like TAVE), although in every other respect they look very similar. So somehow words are influencing the pronunciation of non-words, a finding that can't be explained by the simple dual-route model. Glushko also showed that regularity among neighbours affects the pronunciation of words, which is difficult to explain if skilled readers pronounce words just by directly accessing their lexical entries. GANG is a word with nice friendly regular neighbours ("rang", "sang", "bang" "hang"), but BASE is a word with a nasty inconsistent neighbour ("vase", at least in English English) along with a regular neighbour ("case"). These squabbles among neighbours seem to impede the naming of regular words. Somehow the pronunciation of related words is influencing the pronunciation of regular words, and its difficult to see how this can be explained by the simple dual-route model.

There are several other experiments that show that non-words are not all equal. One of the most revealing was that of Kay and Marcel (1981), who showed that prior exposure to a word could influence the way in which a non-word is pronounced. Suppose you have to read aloud "yead"; in one condition it's preceded by the word "head", and in another by the word "bead". The way in which "yead"

is pronounced is influenced by the pronunciation of the preceding word. It's as though people pronounce the non-word by *analogy* with what comes before. In fact this idea of analogy might have struck you a few pages back when I asked you to pronounce the non-words "smeat" and "nouse"; the latter in particular is very like the word "house", and although introspection is of course highly unreliable, I think that when I pronounce it I distinctly think of the word "house", and model the pronunciation on that. The analogy model of reading came to the fore in the late 1970s when the extent to which there are lexical effects on non-word reading became apparent. The idea is that we read non-words not by grapheme–phoneme conversion, but by finding similar words that we can use as the basis of an analogy (Glushko, 1979; Henderson. 1982; Kay & Marcel, 1981). As we're just reading using words, there's really only one route in this sort of model, and this idea became particularly important later on, as I will very soon show.

The other problem with a simple dual-route model is that it's not obvious how it can handle deep dyslexia. Worse is to come, though, because there are other types of dyslexia that don't fit in either. Patient WLP could pronounce words she didn't understand, a deficit known as *non-semantic reading* (Schwartz et al., 1979). She had great difficulty in retrieving the meaning of written words; she was completely unable to match written animal names to the appropriate picture. She could, however, read aloud their names, and crucially, she was just as good at reading irregular words. So she must be going through the lexicon somehow to retrieve the irregular names, but without accessing the word's meaning.

It's possible to explain lexical effects on non-word reading, neighbourhood effects in word pronunciation, and deep dyslexia and non-semantic reading by making additional assumptions about the model. The dual-route model has to be modified in at least two ways to accommodate all these findings. First, we could explain lexical effects in non-word reading if the non-lexical route had knowledge of spelling–sound correspondences for units larger in size than a phoneme. In particular, we could explain Glushko's results if in addition to GPC rules the route had knowledge of word endings – what psycholinguists call *rimes*. Rimes are the part of the word ending that give rise to rhyme: -ave, -eak, -ouse, -ead, and so on. If the system had ready access to this knowledge it could read easily by analogy. Second, we could explain the dissociations shown between deep dyslexia, where readers can access the appropriate meaning and not the correct sounds, and non-semantic reading, where readers can access the correct sound but not the meaning, if we split the direct, lexical route into two routes. What we would need would be a route that goes directly from print to the lexicon and then to sound without accessing semantics, and another route that gains access to sound through semantics (Patterson & Morton, 1985).

The model is starting to get unwieldy, and has lost the charming simplicity that made it so appealing initially. It's also unclear how the model explains all the symptoms of deep dyslexia, unless we incorporate a Hinton and Shallice type

network into it. A computational model based on the revised dual-route model, called the *dual-route cascaded* (DRC) model, gives some indication of how such a system might work in practice (Coltheart et al., 2001). The model has two core assumptions. The first is that activation is cascaded throughout the network, just as it is in standard connectionist networks; as soon as a unit, such as a letter, is activated, it starts activating those units to which it's connected. The second core assumption is that there's a division of labour between a non-lexical reading system and a lexical reading system, which in turn is split into two routes, one where a representation of spelling is connected directly to a representation of sound, and another that is mediated through semantics. The model can simulate a range of results from experiments on reading, and damaging different parts of the model gives rise to different types of dyslexia.

What's the triangle model of reading?

It looks as though the complexity of the data necessitates a complex model, and compels a division of labour between lexical and non-lexical routes. Connectionist modellers have argued that this division isn't necessary. The most influential connectionist model has become known as the triangle model because of the shape of its overall architecture (Harm & Seidenberg, 2004; Plaut et al., 1996; Seidenberg &

Neighbours, every word needs good neighbours

McClelland, 1989). The triangle model describes a framework for understanding the relation between print, meaning, and sound, comprising orthographic, semantic, and phonological units connected to each other. However, all the work has been on the orthography-to-phonology pathway – print to sound.

The modellers make two strong claims that taken together make the triangle model a radical alternative to traditional models of reading and word recognition. The first bold claim, as you might have deduced from this description of the architecture, is that there is no lexicon in this account. There's no central repository where each word has its own entry and where you access that entry to get at a word's meaning and pronunciation. The sounds of words are patterns of activation across the phonological units; the meanings are patterns of activation across the semantic units; and the print forms are just patterns of activation across the orthographic units. We therefore can't decide if something is a word or non-word by looking it up in our lexicon, because there is no lexicon; all we can go by is what pattern of activation an input string produces across the network. The second bold claim is that we don't have separate lexical and non-lexical reading routes for pronouncing irregular words and pseudowords respectively; instead, the system's statistical knowledge of all spelling–sound correspondences is brought to bear when presented with an input string of letters, and in connectionist models all knowledge is encoded by the weights of the connections between units.

The simulations of orthographic to phonological processing used the sort of architecture you'll now recognise as being very familiar. An input pool of 105 units each representing graphemes (letters) in a particular position in a word (so the /k/ in "cat" was represented by a different unit from the one in "tack") was connected to an output pool of 61 units each representing phonemes (sounds), through an intermediate "hidden layer" containing 100 units. This hidden layer, you will remember, is just necessary computationally for these sorts of models to learn efficiently. Every unit is connected to every unit in the layer above. The model was trained using the back-propagation algorithm on a corpus of almost 3000 monosyllabic words until the model correctly pronounced all of them (a process that takes about 300 iterations – which means the whole process was carried out 300 times – across the whole corpus, or complete sample of words, with more frequent words being presented for training more often). It's important to realise that this corpus contains both regular and irregular words.

What happens when the trained model is then given non-words to pronounce? Plaut et al. presented the fully trained network with over 100 non-words. The network gave the "correct" (by which we mean the same pronunciation as a human would have given the non-word) pronunciation at about the same rate as humans. The important conclusion here is that a single route can produce human-like pronunciations of regular words, irregular words, and pseudowords. How can it do this? Because the connections encode the complete sum of knowledge about spelling–sound correspondences; it's a super-duper analogy model. The network

also gave a good account of other reading phenomena, such as the interaction between regularity and frequency, such that people are markedly slower at reading low-frequency irregular words.

Other simulations explored how surface dyslexia might arise in this architecture, although a full understanding involves broadening the scope of the model. Damage to the orthography–phonology pathway does give rise to reading errors that resemble those of surface dyslexics, but the fit isn't that good. In particular, damage to that pathway led to reading that wasn't bad enough for low-frequency irregular words, and too impaired for regular words, and didn't produce a sufficient number of regularisation errors. These problems led Plaut et al. to explore the possibility that damage to another part of the system leads to surface dyslexia, in particular the idea that surface dyslexic reading reflects the behaviour of an undamaged but isolated orthographic–phonological pathway that has developed with semantic support. After all, children don't learn spelling–sound mappings in isolation; they learn them associated with words that have meaning. So they trained another network where patterns of activation over the phonological units corresponding to words provided some additional support from semantics. After the network was trained, they cut the support from the semantics, and then found that the damaged network performed much more like a surface dyslexic. In terms of the mechanics of the model, what seems to be happening is that as the model learns, there's a division of labour between the semantic pathway, which becomes more responsible for irregular words, and the orthographic–phonological pathway, which becomes specialised in – but not dedicated to – regular spelling–sound correspondences. It's as though meaning glues together phonological patterns that correspond to irregular words. This idea is supported by the observation that with progressive dementia, people usually also become surface dyslexic, presumably because the semantic glue that normally helps us to read irregular words becomes gradually unstuck as the system loses semantic features (Patterson, Graham, & Hodges, 1994).

The triangle model does a good job of explaining normal reading, and explains surface dyslexia in terms of disruption to the semantic–phonology pathway. It would be straightforward to incorporate the Hinton and Shallice (1991) model as part of the orthography–semantic–phonology route. Where does phonological dyslexia fit in? Phonological dyslexia is thought to arise from damage to the representation of phonological information itself, an idea known as the general phonological deficit impairment (Farah et al., 1996; Harm & Seidenberg, 1999, 2001). Clearly words will have much more stable phonological representations than non-words: word sound patterns are much, much more familiar, and they receive support from semantics. So phonological dyslexia occurs when the phonological representations are themselves weakened; phonological representations corresponding to words can always rely on semantic support, but by definition there is no such support available for non-words. The general phonological deficit

hypothesis is supported by the finding that people with phonological dyslexia aren't just bad at non-word reading, but are nearly always also bad at a variety of other tasks involving phonology. For example, they're bad at repeating non-words aloud, and at tasks involving manipulating the sound of words.

So here we have two different explicit accounts of reading: DRC and the triangle. Which, you want to know, is right? Scientists have known for many years that good theories must be falsifiable – that is, capable of being proved wrong. That means they must make predictions. So to distinguish between the dual-route and triangle models we need to be able to identify differential predictions. That is, the triangle model must predict something that could be proved wrong, and the dual-route model something different. There must be something that's right in one model and wrong in another. Unfortunately, life isn't that simple. With models such as these there is always a question about how good is the fit between the real-life data and the predictions of the model; no model purports to capture every aspect of the reading process, so there must be some tolerance in allowing a mismatch between prediction and data, but how much is reasonable? The DRC model is more complex, but its authors claim it accounts for a wider range of phenomena than the triangle model, but while it is good for accounting for the accuracy with which words are read, it isn't so good at accounting for the time it takes us to name words. And both models are limited to the reading of monosyllabic words. Yet again opinions are deeply entrenched and the neutral jury must remain out, but there is an elegance and simplicity to the triangle model that certainly makes it the one to beat (Figure 6.3).

Do we have to sound a word to understand it?

Read the following sentences to yourself, silently and normally:

> Ben planted his seeds in nice tidy little rows. He then got a watering can and watered the rose.

What do you find happens? Do you "hear" an internal voice sound out the words as you speak? What role does this internal voice play? We've spent some time with reading aloud, but most of the reading we do is silent reading. The goal of silent reading is to access the meaning of the words, but must we also access the sound of what we read? There are two ways in which sound might be involved in supposedly silent reading. It would hardly be surprising, given the architecture of the triangle model for example, if semantics activated by orthography then went on to activate phonology incidentally; more interesting is the case that we might have to access phonology to get at meaning, an idea called *phonological mediation*.

There is some experimental evidence for phonological mediation. In the category-decision task, you see a word on a computer screen and have to decide

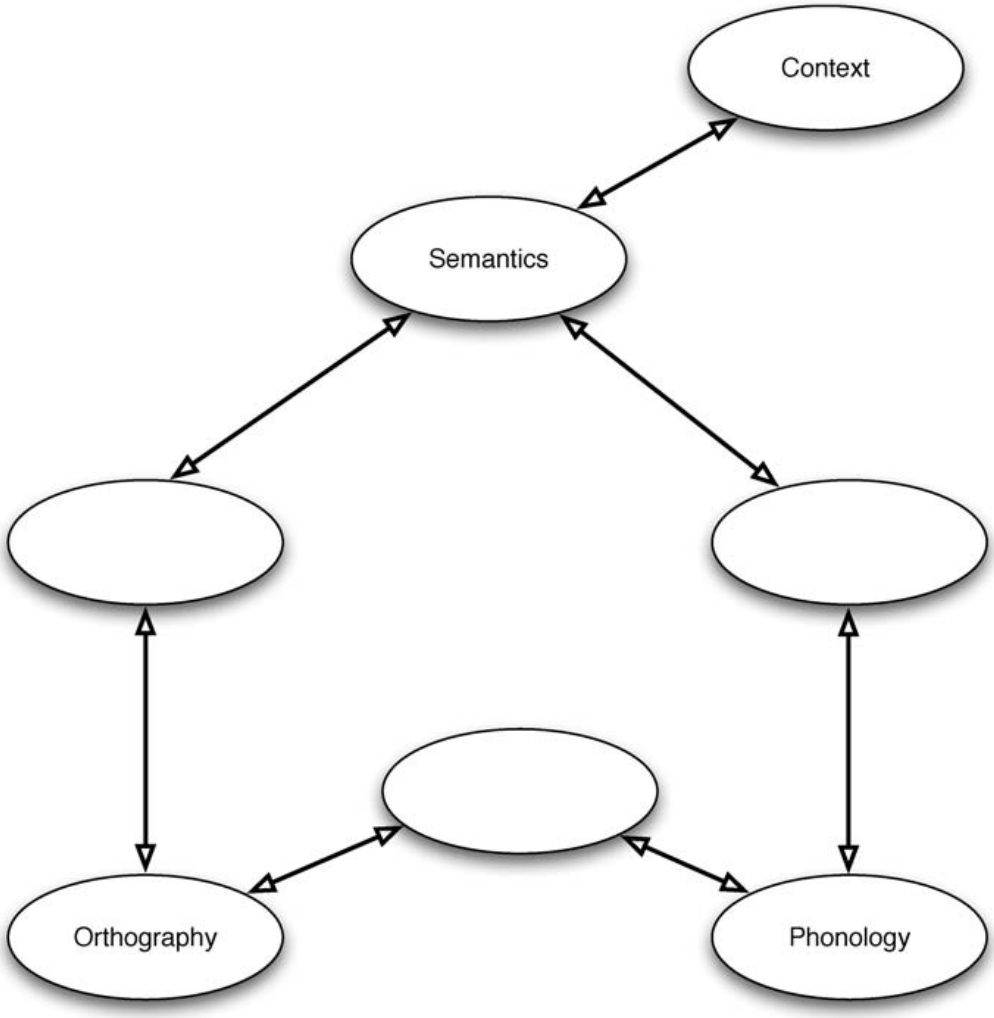

FIGURE 6.3 The triangle model of reading

whether or not the word belongs to a particular category. For example, you might be given the category "fruit", and see the word "pear", in which case you would have to press the YES button. Suppose you see "pair", which is a homophone of a word to which you should respond YES, but to which itself you should respond NO? People make many errors with homophones such as "pair"; it's as though the sound of the word is somehow interfering with the decision-making process (Frost, 1998; Van Orden et al., 1988, 1990). Note that to get this sort of interference the activation of phonology must be relatively early to be able to interfere with the semantic decision-making process.

The idea that we *must* access sound in order to be able to access meaning doesn't really fit in with either of the models we've discussed. The effect does seem to be sensitive to the details of the experiments that demonstrate it; for example, we find more interference if participants have to respond quickly. The effect also

seems to be limited to low-frequency homophones (Jared & Seidenberg, 1991). Furthermore, we only find interference in certain sorts of tasks, such as category-decision. Some dyslexic patients can understand words without being able to read them; they can give perfect definitions of printed words, and match names to pictures, even though they can't pronounce them (Coltheart, 2004). We can explain the results apparently supporting phonological mediation with the idea that although orthography–semantics in the skilled reader is normally the fastest route, in some experimental conditions it might not be. There are two ways to activate the semantic units: the direct orthographic–semantic pathway, and the normally very slow orthography–phonology–semantics (there must be links from phonology to semantics because we can understand words we hear!). So in certain experimental conditions, such as in the category-decision task with low-frequency homophones, there's conflict at the semantic level, as both "pair" and "pear" become activated. It's this conflict that causes the interference and that slows us down and causes us to make errors.

Does speed reading work?

Along with "I used to have the memory of a sieve", advertisements promising faster, more effective reading are common in the Sunday supplements. These advertisements assure you that (for a suitable fee of course) a speed reading course will teach you to read faster. I've seen claims that you can learn to read at the rate of a second a page; if true, this would indeed be a considerable boost, as the normal reading rate is in the range of 200–350 words a minute. If the average page contains about 500 words, that means you can do in a second what it would normally take you well over a minute to do. If these claims are true, I think these courses would be a bargain.

It's true that many people could learn to read a little faster without any cost; sometimes our concentration wanders, sometimes our eyes linger too long on the page. Like all skills, reading skills lie along a continuum, and as with most things most people can probably improve a bit. Sometimes of course we want to read slowly; what's the point of skipping through good poetry as quickly as possible without savouring each word and hearing it in your inner ear?

Unfortunately the psycholinguistic evidence says that you can't increase your reading speed much without cost, and that cost is how much you understand and how much you remember. Just and Carpenter (1987) gave speed readers and normal readers two pieces of text to read – one that was considered "easy" to understand (it was an article from *Reader's Digest*), and another that was "difficult" (an article from *Scientific American*). They then asked both groups some questions to test their comprehension, and found that the normal readers scored 15% higher on both passages than the speed readers. In fact the speed readers

did only slightly better than another group who were asked just to skim through the texts. To be fair, the speed readers still got the general gist of the text, but they were worse at the details. The eyes have it for reading; for a word to be processed properly – which means to access its meaning and to begin the process of integrating that meaning with the rest of the text, it has to "land" close to the fovea, the most sensitive part of the retina, and stay there for enough time for its semantics to become available (Rayner & Pollatsek, 1989). This explanation is supported by Just and Carpenter's additional finding that speed readers couldn't answer questions when the answers were located in parts of the text where the eyes hadn't fixated.

Which isn't to say that you can't learn to improve your memory for and understanding of text. But to do so takes more effort, not less. The most famous method is known as the PQ4R technique (Thomas & Robinson, 1972), and it must work because I've seen it recommended by several famous cognitive psychologists in their textbooks as being the ideal way to read their book. Suppose you need to read a textbook or article and understand it and remember as much of it as possible. First, you *preview* it: you survey the material, look through the contents, find out what's in it, look at the chapter introductions and conclusions, glance at the figures and tables. Speed reading, in the sense of skimming quickly through just to find out what the contents are is useful here. Then you make up some *questions* for each section, trying to make your questions related to your goals for reading the work, turning headers into questions (just as I have already helpfully done for you in this book). Then you should *read* the material carefully, section by section, trying to answer your questions. Next you should reflect on what you've just read; try to relate what you've read to things you already knew; do you really understand what you've just read? If not, why not? Reread the material you don't understand, without worrying too much, because difficult material might take several readings to understand. You might have to look things up elsewhere for clarification. Then after finishing each section you should try to *recall* what you've read; try and phrase it to yourself in your own words. What are the answers to your questions? When you've finished, you should *review* the material, trying to list the main points, arguments, and conclusions. When you've finished, skim through the material again. Repeat an hour later, and tomorrow if possible. And then come back to it several times more. This technique is hardly speed reading! Do I follow this recommended technique? Of course not.

As with everything in life, no sweat, no gain. Having said that, every so often I read about people who read a phenomenal amount. I remember reading that Bill Gates, the founder of Microsoft, often reads two books a day. While I envy him the time he must have available to do all this reading, not to mention the money he has to be able to afford all those books, I do wonder how much he remembers of what he reads.

What is the alphabetic principle?

Learning to talk comes naturally to children; learning to read most certainly does not. Many more children struggle to master reading than have spoken language difficulties. All I remember of my own experience of learning to read is an image of the letters of the alphabet posted around the classroom, with a picture of an apple above the letter "a", and presumably other pictures above the other letters. I have no idea what was above the "Z". Also for some reason great store was placed on being able to repeat the letters of the alphabet in sequence backwards. Somehow I went from reciting the alphabet backwards to being able to read scientific papers and write books such as this one.

I do dimly remember learning to read – in addition to my fun with letters. I remember seeing a picture of a black cat, with CAT printed beneath it, and being taught to spell out "c . . . a . . . t . . ." – "cat"! Easy! But it's not easy. To learn in this way, the child has to know several important things:

1 The spoken word "cat" is made out of three distinct sounds.
2 These sounds correspond to printed letters.
3 The printed word CAT is made up of three distinct elements.
4 And then of course the child has to know that c is the same as C, that a C with serif is the same thing as sanserif, and that my untidy scrawl that looks like a blotch of ink is also a c, and so on.

By "know", I don't mean that the child has to be able to explain this knowledge, just that in some way they have the skill to be able to make use of this information.

The central idea here is that sounds correspond to letters, and this is known as the *alphabetic principle*. Once you know it, you can get a long way. You might never have seen the word COT before, but once you know how the constituent sounds are pronounced, you can spell it out and recognise it. (Let's not worry for now about whether you have to spell it out first in order to recognise it.) Learning and mastering the alphabetic principle is the key achievement of learning to read.

However, children don't just learn the alphabetic principle and go from illiterate to literate; reading development seems to go through a number of phases (Ehri, 1992, 1997). Young children often go through a pre-alphabetic phase, where they haven't yet acquired the alphabetic principle, but can nevertheless still recognise a few words whose shapes they must have learned by rote. So they might recognise the word "yellow" because it's got two tall bits in the middle and a repetitive wiggle at the end. Children at this stage are associating the pattern of a word with the concept, not a particular word. In one famous example a child could recognise the name of the toothpaste, "Crest", but "read" it variously as "toothpaste" on one occasion and "brush teeth" on another. This phase is brief and by no means universal, and is clearly nothing like the direct access of skilled readers, where print

is associated with a word rather than a concept. In the partial alphabetic reading phase, young children have some knowledge of letter names and their correspondences with sounds, particularly the initial and final letters of words. The child is still unable to segment the word's pronunciation into all of its component sounds. In the full alphabetic phase children have full knowledge of letters and sounds and how they correspond, so can read words they've never seen before. Gradually, with more practice, in the consolidated alphabetic phase, children read like adults: words are read directly without the need for grapheme–phoneme conversion, and children are aware that many spelling units, particularly rhymes (often also called *rimes*), are common to many words. Very poor readers will have difficulty getting beyond the third or even second stage.

What is phonological awareness?

To make use of the alphabetic principle you have to know that the spoken word "cat" comprises three sounds – and of course know what those sounds are. This knowledge of sounds and the ability to manipulate them is called *phonological awareness*. There are really two related sorts. You could tell that two words rhymed (horse and course, knight and fight), without being able to decompose and manipulate the sounds. We call this implicit awareness, while the more sophisticated sort of awareness you need to be able to do things with the sounds is called explicit awareness (Gombert, 1992).

Many tasks require phonological awareness; here are a few.

1 What's the first sound of "fun", "doll", and "cat"?
2 What sound do "bat" and "ball", and "cat" and "ham", have in common?
3 How many sounds are there in "cat", "most", and "shelves"?
4 What word would be left if you took the first sound away from "stand"?
5 What would be left if you took the second sound out of "stick"?
6 If you added the sound "t" to the start of "track", what word would you get?

Literacy and phonological awareness are very closely related. Illiterate adults perform poorly on phonological awareness tasks; a group of illiterate adults from an agricultural area of southern Portugal had particular difficulty with tasks manipulating sounds, such as adding and deleting phonemes to and from the start of words; adults who received some literacy training in adulthood performed much better (Morais et al., 1986). What's more, as you might expect, it's literacy in an alphabetic language that matters; adult Chinese speakers literate in both an alphabetic reading system and the logographic Chinese system find these tasks easy, but people who are only literate in the non-alphabetic logographic system find them much more difficult (Read et al., 1986).

Young, preliterate children perform very badly at phonological awareness tasks, and the obvious question is what's the direction of causality here? Does phonological awareness precede literacy, or does literacy in alphabetic languages lead to phonological awareness? Researchers disagree on the answer, although several pieces of evidence suggest that phonological awareness comes first. Training children on phonological awareness tasks leads to an improvement in reading ability (Bradley & Bryant, 1983; Hatcher et al., 1994). Children also seem to be able to be aware of units of speech, such as the beginnings and ends of words, before they learn to read (Goswami & Bryant, 1990). But although all this evidence is suggestive, other researchers urge caution, arguing that these studies failed to take account of the existing level of literacy skills among the children who were tested. Virtually all the studies arguing for a causal role for phonological awareness used children who already had some level of literacy, and by necessity all tasks involving explicit phonological awareness require some instruction or training in the nature of the task, letter names, and the nature of sounds (Castles & Coltheart, 2004).

A related controversy is whether children first learn correspondences between individual letters and the corresponding sounds, or whether they first learn correspondences between larger units and sounds. Young children are certainly aware of syllables from well before they start reading: children as young as four can tap out the number of syllables in words they hear, if instructed clearly (Liberman et al., 1974). One possibility is that children first learn to spot onsets and rimes (Goswami, 1986, 1993). *Onsets* are the beginnings of words: T in "tank" and CH in "church"; *rimes* are the final parts of words, the part that gives them their rhyme, so they'd be -ank and -urch in these two words. The controversy centres on whether the early ability to spot rimes or to segment words into phonemes is the better predictor of early reading ability; as yet there is little agreement (Bryant, 1998; Goswami, 1993; Muter et al., 1998). Although Goswami argues that young children begin reading by analogy, making particular use of the rime, many other studies find that young children need grapheme–phoneme decoding skills before they can learn to read. One telling piece of evidence is that young children find it easier to split words into phonemes than onsets and rimes (Seymour & Evans, 1994). The differences between the results are probably explicable in terms of differences in the materials researchers have used and in the particular instructions given to the children.

What is the best way of learning to read?

My early memory of being taught to read by learning the alphabet backwards can't be right, but learning which sounds go with which letters makes much more sense: "A is for Apple", with a big "A" underneath a picture of an apple; "B is for Bear",

with a big "B" underneath a picture of a bear; "C is for cat", and so on; that method makes much more sense.

The age at which children start to learn to read doesn't seem to make much difference; even if the teaching of reading is delayed until the age of seven, the child soon catches up. The other side of the coin is that very early tuition doesn't seem to provide any particular long-lasting advantage either. There are two different approaches to teaching reading. At one extreme, we could teach the child what individual whole words sound like, an approach called the *whole-word* or *look-and-say method*; at the other extreme, we teach them how to turn letters into sounds, and then how to split words into sounds, an approach called the *phonics method*.

Of course neither method in isolation would work. It would be terribly inefficient to teach every word separately, not pointing out that cat, bat, rat, cats, and rats share something. And learning to spell ghost, aisle, and island by turning the words into their component sounds would lead to disaster. The question is which mechanism do we emphasise most, first. There are several studies that show that the phonics method is greatly superior, and a meta-analysis, which combines the results of several different studies into one, of all the studies carried out on how reading should be taught came to the same conclusion: phonics is better (Adams, 1990; Ehri et al., 2001). Indeed, the key to a child learning to read effectively is their discovery of the alphabetic principle, the idea that letters correspond to particular sounds, and anything that speeds up the discovery of this principle speeds up reading development. Other methods don't work anywhere near as well. Philip Seymour and Leona Elder (1986), then also of Dundee, examined the reading performance of a class of five-year-olds who had been taught without any explanation of the alphabetic principle, and found that these children were limited to being able to read only the words that they'd been taught, and even so they made many errors, resembling dyslexic readers.

Recent work by Rhona Johnston in Clackmannanshire in Scotland caused a flurry of activity in the press when a seven-year longitudinal comparison of children learning to read showed that synthetic phonics is the most effective means of teaching reading. What's more, it confers long-lasting advantages, with children taught by this method still having a reading advantage over children taught by other means several years later (Johnston & Watson, 2007). Synthetic phonics is an accelerated form of phonics instruction that emphasises letter sounds before they are exposed to print. The children are taught a few letters and what these sound like, and they are then taught how these can be blended together in different ways to make up different words. So they might be taught the letters P, T, S, and A, and the sounds /p/, /t/, /s/, and /a/, and then that by combining these they can make the words "tap", "pat", "taps", "pats", and so on. The children are then taught the correspondence between sounds and letters of the alphabet. Children aren't first taught the pronunciation of new words, but have to spell them out from the individual letters.

Teachers emphasise fluency over accuracy. This method enables the basics of reading to be taught in the first few months of the first school year. From 2007, synthetic phonics became the preferred method of teaching reading in the UK.

As Snow and Juel (2005, p. 518) put it, "attention to small units in early reading instruction is helpful for all children, harmful for none, and crucial to some". It's rare in psycholinguistics that we all agree, but this topic is one of those exceptions: reading should be taught in a way that allows the child to discover the alphabetic principle as soon as possible.

I showed in the previous chapter that languages differ in the way in which sounds are translated into print. We saw that even within alphabetic languages there are differences in the regularity with which sounds are mapped on to letters. In Finnish or Serbo-Croat a letter is always sounded in one way, and a sound always corresponds to one particular letter. Put another way, words are always regular in these sorts of language. But in English irregular words abound and the mapping is much more complicated. So we can arrange languages along a continuum of what is called orthographic depth, with Finnish at the shallow end, Greek and German in the middle, and English at the deep end (Goswami, 2008; Seymour, 2005). Languages also differ in the complexity of their syllable structure. In French and other Romance languages the syllable structure is simple and consistent, being mainly consonant–vowel, but in English and other Germanic languages it's more complicated, with a mixture of consonant–vowel (e.g. "ta" in "table") and consonant–vowel–consonant (e.g. "Pat"), and with some syllables involving consonant clusters ("strap" has a cluster of three consonants at the beginning). Given this difference in complexity, it comes as no surprise that French-speaking children acquire the notion of a syllable before English-speaking children (Seymour, 2005). So we can map languages into a two-dimensional grid:

Shallow		Deep
Simple	Finnish, Greek, Italian, Spanish, Portuguese, French	
Complex	German, Norwegian, Icelandic, Dutch, Swedish, Danish, English	

The ease with which children acquire phonological ability and literacy should be greatest at the top left of this grid and most difficult at the bottom right, which seems to be the case. English is *very* difficult, because everything conspires against it. In addition to the complex syllables and deep orthography, there's a very large number of syllables, and the way in which we stress words and sentences is complex and irregular. The English child has eventually to learn whole-word strategies for words like 'yacht" and "cough", rime-analogy strategies for words like "light', "right", and "fight", as well as grapheme–phoneme strategies for regular words (Ziegler & Goswami, 2005). I'm surprised I learned to read at all, let alone write and spell, which causes similar sorts of difficulty.

B is for very, very old banana

What is developmental dyslexia?

In 1896, William Pringle Morgan, a British doctor living in Sussex, described the case of Percy, a 14-year-old boy who, although achieving normally in other academic areas, experienced significant difficulty in learning to read and spell. Pringle Morgan could find no evidence of any brain damage or other obvious explanation, and concluded that Percy was unable to "store visual impressions" of words, what Pringle Morgan called "congenital word blindness".

Learning to read isn't like learning to talk. It isn't that easy, but some children have more difficulty than others. Like most skills, there's a continuum of ability; we'd expect to find good readers and bad readers. Children who find reading very difficult suffer from *developmental dyslexia*. Percy bore all the hallmarks of developmental dyslexia: difficulty in learning to read, difficulty spelling, average performance in other academic areas, and no obvious signs of brain damage or any other obvious impairment that could explain the difficulty.

There's some debate about the nature of developmental dyslexia: is it just very bad reading ability (in which case it's a quantitative difference), or is it something else (in which case there would be a qualitative difference)? Julian Elliott of the

University of Durham received particular attention in the press in 2005 when he questioned the validity of the term "dyslexia" (Elliot & Place, 2004). Although his original claim was that there were so many misunderstandings and misconceptions about "dyslexia" that the term had become virtually useless, this claim became translated, in a way that only journalists can, into "dyslexia doesn't exist". Elliott and others have a point: the term "dyslexia" has become stretched by some people to mean virtually any difficulty in reading, writing, or understanding, including grammatical and memory problems, or indeed, in the extreme, any discrepancy between the child's actual educational attainment and how the parents think the child should perform! It's for reasons such as these that in the past dyslexia was dismissed as "the middle class disease", with the implication that some parents don't like to accept that their child might not be excelling academically for other reasons.

If we wished to be pedantic, there's another possible confusion, in that developmental dyslexia is a difficulty with reading. Difficulty in learning to spell and write properly is called developmental dysgraphia. Unfortunately, developmental dysgraphia and dyslexia always go together, so we can use "dyslexia" to refer to both. All of this controversy and confusion is a great pity because of course developmental dyslexia exists, but the definition I'm going to work with is quite strict: it's a marked discrepancy in reading ability skills and non-verbal measures of IQ. The different ways in which the term "dyslexia" is used, and how it's measured, make its prevalence in the population difficult to estimate, but it's likely to lie between 10% and 4%. Note that developmental dyslexia affects someone all their lives, so although more attention has recently been paid to diagnosing it in childhood, there must be hundreds of thousands of adults in the UK who have undiagnosed developmental dyslexia.

There are different types of acquired reading disorder, so the question naturally arises as to whether or not there are different types of developmental disorder. Castles and Coltheart (1993) identified two subtypes of developmental dyslexia. People with surface developmental dyslexia have particular difficulty with irregular words. Their reading of non-words and their ability to convert graphemes into phonemes is relatively good, but they have difficulty in constructing the direct-access reading route. (I'll put aside the question of whether or not there are direct and indirect routes for now and put it this way for clarity, but we'll come back to an alternative interpretation shortly.) People with phonological developmental dyslexia have difficulty with grapheme–phoneme conversion; they have difficulty with non-words and spelling out new words, but can read both regular and irregular words they know by the whole-word, direct-access route. This division into good and bad grapheme–phoneme conversion, phonological skills versus good and whole-word, orthographic skills, is not limited to dyslexia. Adults in the population within the normal range of reading skills also vary in the degree to which they rely on these skills. Baron and Strawson (1976) distinguished between what they called

"Chinese readers", who are relatively good at orthographic skills but relatively poor at phonological, and "Phoenicians", who are relatively good at phonological skills and bad at orthographic (and of course some individuals will be good at both). Hence within the normal population we can identify individuals with a very mild form of phonological dyslexia and others with a very mild form of surface dyslexia.

What causes developmental dyslexia?

Researchers have proposed several causes of developmental dyslexia. From the outset I should make clear that there might be more than one type and more than one cause of developmental dyslexia, particularly given that there's more than one type of deficit. We should also bear in mind that there are also different levels of explanation, so that a genetic impairment might lead to a child having difficulty in representing sounds, for example.

Although I've been at pains to point out that dyslexia is a selective deficit in reading and spelling relative to other non-verbal skills, the reading problem does seem to be associated with a cluster of other problems. Sufferers tend to suffer more than we would expect from a depressingly lengthy list of additional problems, including dyspraxia (difficulty in planning movements), clumsiness, difficulty in processing sounds, difficulty in producing neat handwriting, and difficulty in producing sounds. None of these will make the dyslexic child's life any easier, but the extent to which they are necessarily related to pure dyslexia is uncertain. There is strong evidence that developmental dyslexia runs in families, and some researchers have tentatively identified a number of regions on chromosomes that might be implicated in dyslexia (Fisher et al., 1999; Pennington & Lefly, 2001; Schumacher et al., 2007). A genetic cause of, or at least predisposition to, dyslexia makes perfect sense, and it is also very plausible that it might be associated with this cluster of additional symptoms, although the precise way in which all these things hang together remains to be worked out.

In the popular mind at least, one of the most well-known theories of the origin of dyslexia is that it arises because the person has difficulty in visual perception. They tend to confuse similar looking letters, such as p and b, and m and n, and often report that the letters appear to dance across the page. These observations could be explained if the dyslexics have difficulty keeping the eyes stable when fixating on print, or difficulty in resolving fine detail in print, or both (Lovegrove et al., 1986). Brain imaging studies show increased activity in the occipital region of dyslexics, that part of the brain where much low-level visual processing is carried out, presumably reflecting the extra work necessary to try to make sense of print (Casey et al., 2001). Another low-level account says that dyslexia arises from a deficit in auditory processing. Dyslexics have difficulty processing rapidly changing sounds, of which speech is a perfect example (Wright et al., 1997). Yet another

account says that dyslexics have a mild impairment to the cerebellum, the region of the brain responsible for automatic processes, co-ordinating perception and action, and motor control and co-ordination (Nicolson et al., 2001). The cerebellum is the little bump at the bottom of the brain, just behind the midbrain where the spinal cord enters the brain (look back at Figure 1.3). Brain imaging suggests that there are processing differences between the cerebellums of normal and dyslexic people (Ramus et al., 2003).

A theory that attempts to bring together the visual and auditory and other deficits is that people with dyslexia have an impairment to the *magnocellular* perceptual pathways in the brain (Stein, 2001, 2003). These pathways, characterised by large cells, respond quickly to contrast, movement, and rapidly varying stimuli; although most clearly understood for visual processing, there seems to be an analogue in the auditory processing system. Furthermore these pathways have particularly string inputs to the cerebellum. So here we have a candidate theory that can bring together many of the phenomena and observations of dyslexia: a genetic abnormality leads to dysfunction of the magnocellular pathways, leading to difficulties in processing rapidly changing visual and auditory processing, as well as abnormal development of the cerebellum leading to difficulties in motor control. While the magnocellular theory provides an impressive integration of a range of data, it's unlikely to be the only explanation of developmental dyslexia because not all people with dyslexia have any obvious visual processing deficit (Lovegrove et al., 1986), and some people with dysfunction of this pathway do not develop dyslexia (Skoyles & Skottun, 2004).

Another explanation at the biological level focuses on the role of the *planum temporale*, a structure that lies at the heart of Wernicke's area, that region of the left hemisphere of the cortex known to play an essential role in processing language. Although the planum temporale is usually significantly larger in the left hemisphere than in the right (up to five times larger, making it the most asymmetric region of the brain), this asymmetry is much less in people with dyslexia (Beaton, 1997). Damage to the planum temporale is associated with difficulties in phonological (sound) processing, and in particular difficulties in processing sounds in real time. Autopsies of four (dead) people with dyslexia found this asymmetry, but also showed abnormalities of the structure consistent with abnormalities of the migration phase of foetal development, when developing neurons move to their final destination in the brain (Galaburda et al., 1985). Again the line of argument is that the reading deficit in developmental dyslexia follows from a primary deficit in processing sounds.

Although Castles and Coltheart argued that there were two distinct subtypes of developmental dyslexia, surface and phonological, most researchers conclude that dyslexics fall on a continuum, with the surface type at one extreme and the phonological at the other (Manis et al., 1996; Wilding, 1990). Those near the surface end have difficulty with irregular words but are not so troubled by non-words and

grapheme–phoneme conversion; those nearer the phonological extreme will show the opposite pattern. Those individuals near the middle will have both sorts of difficulty. Children at the surface extreme perform very similarly on a range of tasks to reading-age-matched control children; that is, they "just" seem to have difficulty reading irregular words. It's as though, for some reason, they're learning to read much more slowly than normal – they're delayed readers. Children with signs of phonological dyslexia show other deficits relative to reading-age-matched controls, however: they're impaired at a range of tasks needing phonological skills, including phonological awareness, non-word reading, picking out phonologically distinct words, and reduced short-term memory (e.g. Bradley & Bryant, 1983; Campbell & Butterworth, 1985; Goswami & Bryant, 1990; Metsala et al., 1998).

The idea then is that developmental phonological dyslexia arises from a general phonological deficit, in just the same way as in the one-route triangle model acquired phonological dyslexia arises from a general phonological deficit. Exactly how this system might work is demonstrated in a connectionist model of reading (Harm & Seidenberg, 1999, 2001). These authors showed how the symptoms of developmental phonological dyslexia are generated by damage to the phonological units representing sounds before the model is trained to read. The symptoms of developmental surface dyslexia could be generated in several ways, such as providing less training, making training less effective, and degrading the visual input to the network (resembling a real-life visual impairment). They also showed that it's possible to have a phonological impairment that's severe enough to cause massive disruption to learning to read, but at the same time isn't severe enough to interfere with speech perception or production.

And that, I reckon, is what's wrong with me: I have a mild general phonological deficit. I have difficult in learning new words, manipulating sounds, repeating and remembering non-words, learning foreign languages, needed speech therapy as a child, and all my life have always made many, many speech errors – all characteristics of such a problem. I can get by very nicely, thank you, because of very strong semantic support to words.

Here then we have an account that potentially synthesises all the above. We seem to have two types of biological deficit, one involving some disturbance of visual processing and one involving a disruption of phonological processing, probably arising from some genetic disorder. We have a computational model that shows how a continuum of surface and phonological dyslexia can arise from delayed reading and a general phonological deficit.

The remaining obvious question is how developmental dyslexia should be treated. The first point to make is that just because there's a genetic predisposition doesn't mean that a person will develop full-blown dyslexia. In "at risk" families, scores on measures of phonological skill lie on a continuum; for young children, the higher the score, the greater the likelihood that they will later develop dyslexia

(Pennington & Lefly, 2001; Snowling et al., 2003). What's more, good general language skills, particularly good early vocabulary development, can partly offset the phonological deficit. Nevertheless, the deficit will still be apparent with phonological awareness tasks, and even children classified as normal readers will have some difficulty reading and spelling non-words. So in a family at risk, there might be widespread subtle language difficulties, but other skills can compensate. So one thing that might help would be providing additional training from an early age if children at risk can be identified. Generally training on tasks that introduce and improve phonological awareness from as early as possible is desirable (Bradley & Bryant, 1983; Snowling, 2000). For example, children who were poor at rhyme judgement were trained individually and weekly for two years at tasks such as having to group "hat" with "cat" on the basis of rhyme, but with "hen" on the basis of the initial sound. After four years the experimental group performed significantly better at reading and spelling than the untrained control group. Providing extra training on a range of phonological skills is without doubt the best way to improve the reading ability of young children with developmental dyslexia (particularly those towards the more phonological end of the continuum), but is also the best way to improve the skills of poor readers in general (Hatcher et al., 1994). Training can be advantageous beyond the earliest years: DF, a 10-year-old boy with poor reading ability characterised by surface dyslexia, showed a marked improvement after training on many low-frequency irregular words, while SP, an 11-year-old boy with poor reading characterised by phonological dyslexia, showed improvement after training on phonological awareness skills (Broom & Doctor, 1995a, 1995b).

For children at the surface end of the continuum, training on irregular words, particularly low-frequency words, is a great help. For those whose dyslexia arises because of a visual impairment, improving the child's magnocellular system by training eye fixations has been shown to help in at least some cases (Stein, 2003). Other methods of improving visual clarity and reducing interference might work; there is some anecdotal evidence that orange or yellow paper or overlays or tinted glasses might assist some people, although the experimental evidence is currently a bit scanty (Stein, 2003; Wilkins, 2003; Wilkins & Neary, 1991).

There's no doubt that people with dyslexia are disadvantaged in educational fields requiring literacy – which to some extent is most of them. In the UK it is quite rightly illegal to discriminate against people with dyslexia, and educationalists and employers must make reasonable adjustments. The question naturally arises as to what is a reasonable adjustment.

We do people with dyslexia no favours at all by over-compensating, or by giving them useless or incorrect treatments or adjustments. In some institutions students might be given 15 minutes' extra "reading time" for a three-hour exam. What's the reasoning behind this? It's reasonable for complex exams involving a great deal of reading, but for short exam questions, how will it help at all? For a

dyslexic person, the disorder is mainly going to show itself with difficulty in writing and in spelling correctly. How will extra time reading help that? What would help the person most would be help in transcribing, or taking spelling difficulty into account when marking. Similarly instructions given to educationalists are very vague. Stickers on work saying "This person is dyslexic; please take this into account" are common, but how should the marker most fairly take this into account? By ignoring spelling mistakes? More clarity is needed, for the sake of both the dyslexic person, who might not receive the help they need, and the non-dyslexic person, who could end up finding themselves disadvantaged.

How do we understand ambiguous words?

What do the following utterances have in common?

I'm going to the bank.
What a lovely pen!
That's some ball.

Apart from being not terribly interesting, they're all *ambiguous* sentences – they all have more than one meaning. In the first example, we could be off to take out our life savings from the money bank, or for a nice quiet day's fishing on the river bank. In the second, I could be complimenting you on your nice biro, or on your cunningly made area for holding sheep. In the third, I could be praising the little round thing the children are kicking around the park, or the dance extravaganza I'm gazing down at from the balcony. These sentences are ambiguous because the words "bank", "pen", and "ball" (and many others) are ambiguous – they have multiple meanings, or *senses*.

These are words ambiguous both in writing and sound, but some words are ambiguous just when we hear them. These words are called homophones.

What a lovely night!

When we hear this sentence, we could interpret it to mean the night is very pleasant, or the knight is a darned good-looking chap.

And to complicate things even more, some words can belong to two different syntactic categories:

Hold on, the plane is going to bank suddenly.

This type of ambiguity, where words have multiple senses, is called lexical ambiguity. In the next chapter I'll introduce another sort of ambiguity, syntactic

ambiguity, which occurs when syntactic structures can have more than one interpretation.

Lexical ambiguity is one of the most important sources for humour, particularly most puns, or plays on words. Type "best puns in the world" into your favourite internet search engine and you will find such gems as:

> Two silkworms had a race. They ended up in a tie.

This joke gets its humour (such as it is) from the lexical ambiguity of "tie", which can refer either to an identical position in a race or game, or an item of clothing. Or you could look at some clips from the best James Bond puns ("I see you handle your weapon well", which gets its humour from . . . well, you get the idea). You could also try typing "ten worst puns in the world" into Google, and you'll discover that some of the same puns are in both the best ten and worst ten puns of all time. Oh well.

In terms of our model of language processing so far, the same word – or, put more precisely, the same pattern of activation over phonological or orthographic units – is pointing to different patterns of activation across the semantic features, or two different semantic attractors. Yet we rarely make a mistake and end up with the wrong meaning; most of the time we're not even aware that a word is ambiguous. How do we so easily end up with the right meaning?

It's obvious that we use the context around the ambiguous word to latch quickly on to the right meaning. Given:

> I'm going to take some cheques to the bank.
> The fisherman put his rod down on the river bank.
> My pencil's broken – could I borrow your pen?
> The farmer rounded up the sheep into the pen.

I doubt that without the preceding preamble you'd even notice the ambiguity. One of the earliest models of how we cope with lexical ambiguity, the *context-guided single-reading lexical access model* (e.g. Schvaneveldt et al., 1976; Simpson, 1981), said that context somehow restricts the access process so that usually only the correct meaning is accessed. The problem with this model is in that word "somehow"; at the time it wasn't understood how context could have such immediate and powerful effects. The context of an utterance is a big thing; it isn't just the other words in the sentence, it's what else you've been talking about in the conversation, what you're looking at, what you've been looking at, all knowledge of the word – context is huge, and potentially any of it could be brought to bear on resolving an instance of lexical ambiguity, while how context works is actually very mysterious.

An alternative early model was the *ordered-access model* (Hogaboam &

Perfetti, 1975). You'll probably have noticed that the different senses of most ambiguous words aren't equally commonly intended. "Pen" has a high-frequency sense (writing instrument) and a low-frequency sense (animal enclosure), as does "bank" (financial institution and river side, although the difference in meaning frequencies isn't as pronounced). There is the complication that frequency is a personal thing: fishermen and financiers will see things differently. Let's talk about the average frequency of sense across people. According to the ordered-access model, when we come across an ambiguous word we check the most frequent sense first against the context to see if it makes sense, and if it does not, then we move on down to the next most frequent sense, and check that.

Early experimenters lacked sufficiently sensitive techniques to make much headway in distinguishing between these models. The breakthrough came with the development of the technique called *cross-modal priming*. In tasks based on this technique participants have to juggle two things at once: they have to listen and watch. In Swinney's (1979) experiment, participants heard little stories such as:

> Rumour had it that, for years, the government building had been plagued with problems. The man was not surprised when he found (several spiders, roaches, and other) bugs1 in the cor2ner of his room.

I'll come back to that 1 and 2 later. The ambiguous word here is of course "bugs". Half the participants heard a version which included the strongly biasing context "several spiders, roaches, and other" which pushes you to one sense of bugs; the other half just heard "found bugs", without any strongly biasing context.

At the same time participants saw words or non-words on a screen in front of them to which they had to make a lexical decision, so they had to press one button if they saw a word, and another if they saw a non-word. Swinney measured the time it took people to make this decision. I'll only talk about what happened when they saw words; in this experiment some non-words are slipped in just to keep participants on their toes so they really do have to read the word and can't keep pressing "yes" all the time. The targets were "ant" (which is a word associated with the biased sense of the word), "spy" (a word associated with the irrelevant sense in the biased context), and "sew" (a neutral word that provides a baseline).

Remember that semantic priming is a very robust effect: it's easier to recognise a word if we've just been exposed to one related in meaning. So if both senses of "bug" are active, people should be faster to respond to both "ant" and "spy" than the baseline, "sew", but if only the biased sense is active, then they should only be faster to respond to "ant", with the irrelevant sense, "spy", being the same as "sew". Swinney found that the pattern of results depended on the timing. Immediately after the ambiguous word "bug" (indicated by the 1 in the example above), both senses were active (both "ant" and "spy" were facilitated), but very soon after, at point 2, after just a few intervening syllables, only the relevant sense, "ant", was

facilitated. This result shows that when we come across an ambiguous word, all the senses are activated immediately, but the activation of the irrelevant ones dies away very quickly. A similar experiment measuring how long it takes people to name a word rather than make a lexical decision came to a similar conclusion (Tanenhaus et al., 1979). So we access all the senses first of all, but use the context to decide between them very quickly – within 200 milliseconds.