Chapter 5

Meaning

chapter 5

AT SCHOOL THERE WAS a new teacher who made the dreadful mistake of starting off nice. You can't give schoolboys an inch. (I assume he's learnt his lesson and is now nasty and very successful somewhere.) Very soon his classes degenerated into mayhem. One day the headmaster walked into the riot and demanded "What is the meaning of this?" We didn't know how to answer him then, and I still don't really know what he meant. There – without noticing it, I referred to "meaning" again.

"Meaning" is one of those many words in language that we think we know the meaning of, but the more we think about them the more slippery they become. Meaning underlies language – it's the starting point of language production and the end point of language comprehension. It underpins all activities: linguistic, cognitive, and social. Without it our lives are unthinkable – dare I say it, meaningless?

What has meaning? Words do. Sentences do. The meaning of sentences is derived from the meaning of the constituent words interplaying with the syntax. But do objects have a meaning? Does a conversation? Some even talk about the meaning of life. At least if psycholinguists aren't sure what meaning is, they have happily given its study a name: semantics.

What's the meaning of meaning?

I had a friend at University who mocked philosophy and philosophers mercilessly. Whenever either was mentioned he would screw his face up and say in a high-pitched whine, "Yes, but what's the meaning of meaning?"

As ever, when stuck for meaning, let's look it up in a dictionary. And, as ever, let's be prepared to be disappointed. The online Freedictionary says meaning is:

1 something that is conveyed or signified; sense or significance
2 something that one wishes to convey
3 an interpreted goal, intent, or end
4 inner significance.

Here the act of looking "meaning" up in the dictionary is perhaps more interesting than what we found. Many people find out the meaning of something by looking it up in the dictionary; therefore, meaning is the dictionary definition. Perhaps we each have a mental dictionary, and a word's meaning is what's stored in its entry? And when we learn a new word, we create a new entry? And the meaning of a sentence somehow combines all these entries, indicating how they are related?

The problem with the dictionary account is that in the end we just go round in circles. By definition, words have to be defined in terms of other words. Pick a word at random in a dictionary. "Meaning" will do. Part of its definition is the word "sense". Look that up, and, among other things, we get:

A meaning that's conveyed.

That was a pretty small circle! Nearly as short as Steven Pinker's (2007) example:

Endless loop: n. See loop, endless
Loop, endless: n. See endless loop.

If we had an English-speaking man who knew no Chinese, and who sat in a little room with a big Chinese dictionary, and we gave him Chinese words written on scraps of paper that we fed him under the door, would we wish to say that this man knew the meanings of these words? Of course not. What this account clearly lacks is any reference to the outside world. Bringing the outside world into psycholinguistics and the brain is one of the big challenges facing modern psychology, and has been surprisingly largely ignored by psycholinguistics until recent times. No wonder we've gone round in circles.

What's a dog?

My friend was wrong about philosophers (and linguists). One very useful distinction they've given us is between the denotation and connotation of the meaning. The *denotation* is the primary meaning – its core, essential meaning that everyone agrees on (or would agree on if they expressed it). There should be no room for doubt about the denotation of a word (although one sometimes wonders, as when Bill Clinton said "it depends upon what the meaning of the word 'is' is"). When I talk about "semantics" I mean denotation. The *connotation* is the secondary meaning – all the associations we have to a word.

We can see that meaning somehow has to refer to the world; the real problem is what is the nature of this relation. For some things it seems very easy. There's just one moon (pedants, please, I'm just talking about our moon, the one everyone talks about) and just one sun, so the meaning of "moon" and "sun" relate in quite a straightforward way to the world. With "dog" things get a little trickier, because there's a very large number of them, of many different sorts, but at least they're very obvious things that form a pretty straightforward category (we call them a *natural kind*). "Mammal" gets trickier still, but we can still use the very useful following trick: we can point to something and say "yes, that's a mammal", or "no, it's not". Perhaps the meaning is in some way related to the rule that enables us to decide whether or not something is a member of a category? But then we get on to yet more abstract words, such as "truth" and "justice"; it's obvious that these words don't refer to objects, so our pointing rule won't work any more. We can apply the same sort of analysis to adjectives: we can point to things that are yellow or dead, admitting that we might be troubled by relativity (some things might be big relative to others, but relative to other things they might be small) or subjectivity (beauty is in the eye of the beholder, they say). We could do something similar for verbs, but with more difficulty. Grammatical words – let's not even go there yet.

Referring to things or properties in the world, then, is an important aspect of meaning. Our knowledge of meaning enables us to say whether a thing is a dog or not a dog, but this is most definitely not the same thing as saying that we represent the meaning of a word like "dog" as some kind of decision rule. The psychological representation of meaning is sufficiently powerful to enable us to do this sort of thing, but that doesn't mean it is that thing.

And here's another problem with pointing: we don't always know what things are in the world. The classic example is that of Hesperus and Phosphorus. The early Greeks called the bright star that sometimes lights up the evening western sky "Hesperus"; they called the bright star that could be seen some mornings in the east "Phosphorus". We now know that they both refer to the same thing – the planet Venus. Both words refer to the same thing, but it doesn't seem at all right to say therefore that for the ancient Greeks they had the same meaning. We need to distinguish two aspects of meaning: the thing referred to in the world, and the sense

that captures the world as we understand it. The *intension*, or sense, is our internal, abstract specification that enables us to pick things out in the world; the *extension*, or reference, is the thing, or set of things, referred to. So for the Greeks Hesperus and Phosphorus had different intensions but the same extension.

This debate isn't as philosophical as it might at first seem because our knowledge of the world changes. A little while ago Pluto was demoted from being a planet to a dwarf planet. It's still the same thing (it still has the same extension), but its intension has changed.

What is a dog, then? I phoned up my mother and asked her that question. Being apprehensive about what psychologists get up to, I think she suspected it was a trick question, and needed some reassurance before she said "An animal with four legs. It's kept as a pet. I could go on". I then asked the mother-in-law too, and got: "A small domesticated wolf, often referred to as 'man's best friend' because of its loyalty to its owner". Neither is terribly good in my opinion! (I hope they don't read this.) Yet I'm sure they know a dog when they see one. In this case at least our implicit knowledge of meaning is demonstrably better than our explicit knowledge. But I then started getting carried away with people's definitions of "dog", so I asked a friendly professional psycholinguist, and got: "I'd look in a dictionary". Undeterred, I asked another and got: "Animal who walks on four legs, has a tail that wags in response to happiness or interest in surroundings, can be domesticated; prefers to live in packs (but not necessarily with other dogs, but could be with people, for example), lives in the moment, loves to play, fiercely loyal to the pack and will defend said pack courageously". I should have known better than to ask psycholinguists. And I thought it was just a four-legged animal that barks and is kept as a pet.

Let's look at it the other way round, and when we do, we see that we have a preferred way of talking about things. Let's try another experiment. If you show people the picture on p. 122, and ask them what it is, what are they most likely to say?

Most of them will say "dog". They're unlikely to say "animal", very unlikely I think to say "mammal", and probably unlikely to give the breed of dog. "Dog" seems to be just the right level of specificity for talking about things most of the time; it's got the right balance between being just informative and discriminative enough (in a way that "animal" and "mammal" aren't), on one hand, and general and economical on the other. This sort of level ("dog", "cat", "chair", "car") is called the *basic level* (Rosch, 1973, 1978). Above the basic level we have one or more superordinate categories ("mammal", "animal", "furniture", "vehicle"), and beneath we have category members or subordinates ("poodle", "Siamese", "office", "easy", Volkswagen, Chrysler), which in turn might be subdivided. We tend to categorise, and perhaps think, at the basic level; there's a large loss in distinctiveness as we go from the basic level to the superordinate category, but not much to be gained most of the time by making unnecessarily fine distinctions between subordinates below. Objects at the basic level tend to look alike, at least in profile. Basic-level objects have several psychological advantages: children usually

learn basic-level names first, it's the highest level at which we can form a mental image (try forming a mental picture of "animal" without thinking of something more specific), people can find most things to say about basic-level things, and we process basic-level names more quickly than those at other levels (Jolicoeur et al., 1984; Rosch et al., 1976). When people misremember stories, they move from the direction of subordinate terms to using the basic-level name (Pansky & Koriat, 2004). Of course the basic level might change depending on one's level of expertise; my knowledge of seagulls used to be such that the basic level for me was the species, rather than seagull, or bird.

None of this is to deny that some categories are fuzzy (in the way that the distinction between this chapter and the next is now getting fuzzy). Category membership is determined by the underlying concept, but that's straying more deeply into the realm of meaning.

Are meanings captured by networks?

A dog's an animal. A setter and a poodle are sorts of dogs. Everything that's true of an animal is also true of a dog, and everything that's true of a dog is also true of

Another artistic composition, this one of Hesperus. It's there, honest

Mum. What's this? It's my friend Felix

setters, poodles, Alsations, and rottweilers. Collins and Quillian (1969) realised that once you've specified information at one level, you don't need to do it again at a lower one. They presented a model of semantic memory known as a semantic network. In their model, knowledge is stored in a hierarchy. Concepts are represented as nodes connected by links; these nodes are connected by links, and these links can have values. In their model, the most common one is called an ISA link; no prizes for guessing what ISA means ("is a" or if you want to pad it out a bit, "is an example of"). Attributes are stored at the highest possible node; so "has wings" is stored at the "bird" node, because it's true of all birds, but not all animals. It's easy to understand when seen in a diagram.

Models are fine, but are much more impressive when supported by experimental evidence. How can we test this sort of hierarchical model? Think about how we might verify a statement such as "a canary is a bird". We'd start off at the CANARY node, and travel up to the BIRD node. The two nodes are connected by an ISA link, so the statement is true. What about "a canary is an animal"? Just the same, only this time we have longer to travel to establish a connection. So the crucial prediction is that it should take longer to verify "a canary is an animal" than "a canary is a bird". What about "a canary is yellow"? Easy, we retrieve the IS YELLOW information directly from the CANARY node, so we should be relatively fast. What about "a canary has wings"? We have to travel up to the BIRD node to be able to retrieve

that information, so that should take longer. And for "a canary has a liver",we have to go all the way up to the ANIMAL node, which should take longer still.

Collins and Quillian tested these predictions using a *sentence verification task*. This task is very simple: you present people with a sentence on a computer screen, such as "a canary has wings", and ask them to press one key if the statement is true and another if it's false, and you measure how long it takes them to make the decision. Obviously it's going to take people some time to read the sentences and some time to press the key, but these should be constant across sentences. Any *differences* in response time should therefore reflect differences in decision time.

The results from the sentence verification task supported the model. Essentially the further you have to travel along the hierarchy, the longer it takes you to retrieve that information and make a judgement about the veracity of a sentence containing it. We have to make the additional assumption about how people decide on the falsity of statements ("a fish has wings") somehow, perhaps by going up until we find a superordinate that can then lead us down to the appropriate node again (we can go up from fish to animal and then down another branch to find bird, which has wings), or perhaps just by rejecting statements as false if we don't find a match quickly enough.

It's been worth spending some time with this early model because although no one now thinks this is how we store all information about word meaning, the basic idea that meaning is represented by the interconnection of concepts is still very much alive. Some of the problems with the model are obvious: it's all very well for natural kind terms like birds and canaries, but what about our old friends truth and justice? Where do they live on a network? And think about "canary" and "wings" compared with "canary" and "liver"; the first two are always turning up together in sentences – we can say they're highly *associated* – but "canary" and "liver" do so much less frequently (I've never seen them together on the same page before now, I think) – they have a very weak associative strength. When we control for the strength of association of the words in the sentence, the linear distance effect is weaker, but not eliminated (Conrad, 1972; Wilkins, 1971). Other experimental results followed, showing that verification time isn't related to cognitive distance. We're faster to verify "a cow is an animal" than "a cow is a mammal", even though mammal is closer to cow than animal (Rips et al., 1973). If two things are related in some way, we find a false sentence more difficult to reject than one in which the two things are unrelated:

A pine is a church.
A pine is a flower.

Both are false, but the relatedness of the two words leads us to find the second sentence harder to reject, and therefore we're slower (Wilkins, 1971). And we're faster at dealing with items that are more typical of their category than ones that are not, a result called the typicality effect:

A penguin is a bird.
A robin is a bird.

Both are true, but robins are in some way "better" birds than penguins – they're more typical birds – so we're faster at verifying the first sentence than the second (Figure 5.1) (Rosch, 1973).

There are of course some pretty obvious modifications we could make to the model. Perhaps the most obvious is that although we called the model a network, it isn't really. We could introduce more links between things and vary the lengths of the connections, so that, for example, the ROBIN node is closer to the BIRD node than is the PENGUIN node. We haven't really talked about the mechanism whereby verification occurs – *what* travels along the links? If we think back to the first chapter, an obvious candidate is activation, that mental energy that inhabits our mental networks. The longer the link, the longer the activation takes to get there. Activation spreads out from an activated node to all those connected to it. These modifications constitute the Collins and Loftus (1975) *spreading activation network* model of semantic processing.

The details of these models needn't detain us. We've established a very

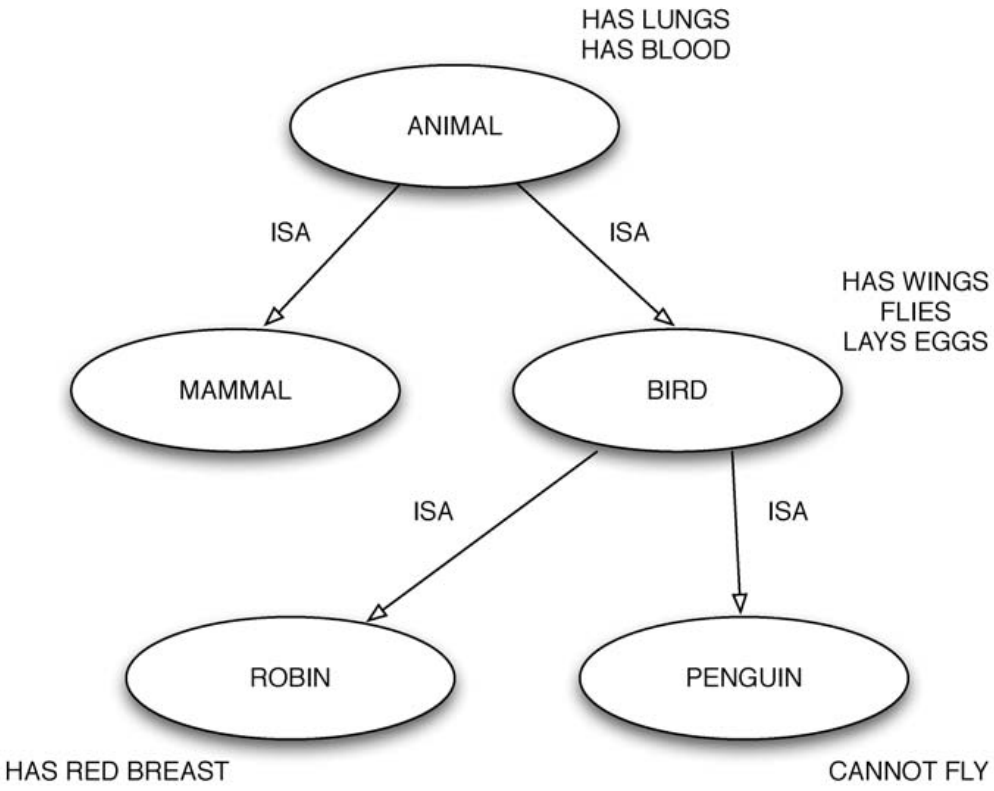


FIGURE 5.1 A small part of a Collins and Quillian type semantic network, showing property inheritance

important principle: knowledge can be represented in an interconnected network of information through which activation spreads (Figure 5.2).

What's a semantic feature?

Whatever their numerous and manifest deficiencies, the definitions I've discussed so far do have one subtle but important feature in common: they try to explain the meaning of something in terms of combinations of simpler units of meaning. The Collins and Quillian model does this to some extent as well: information such as

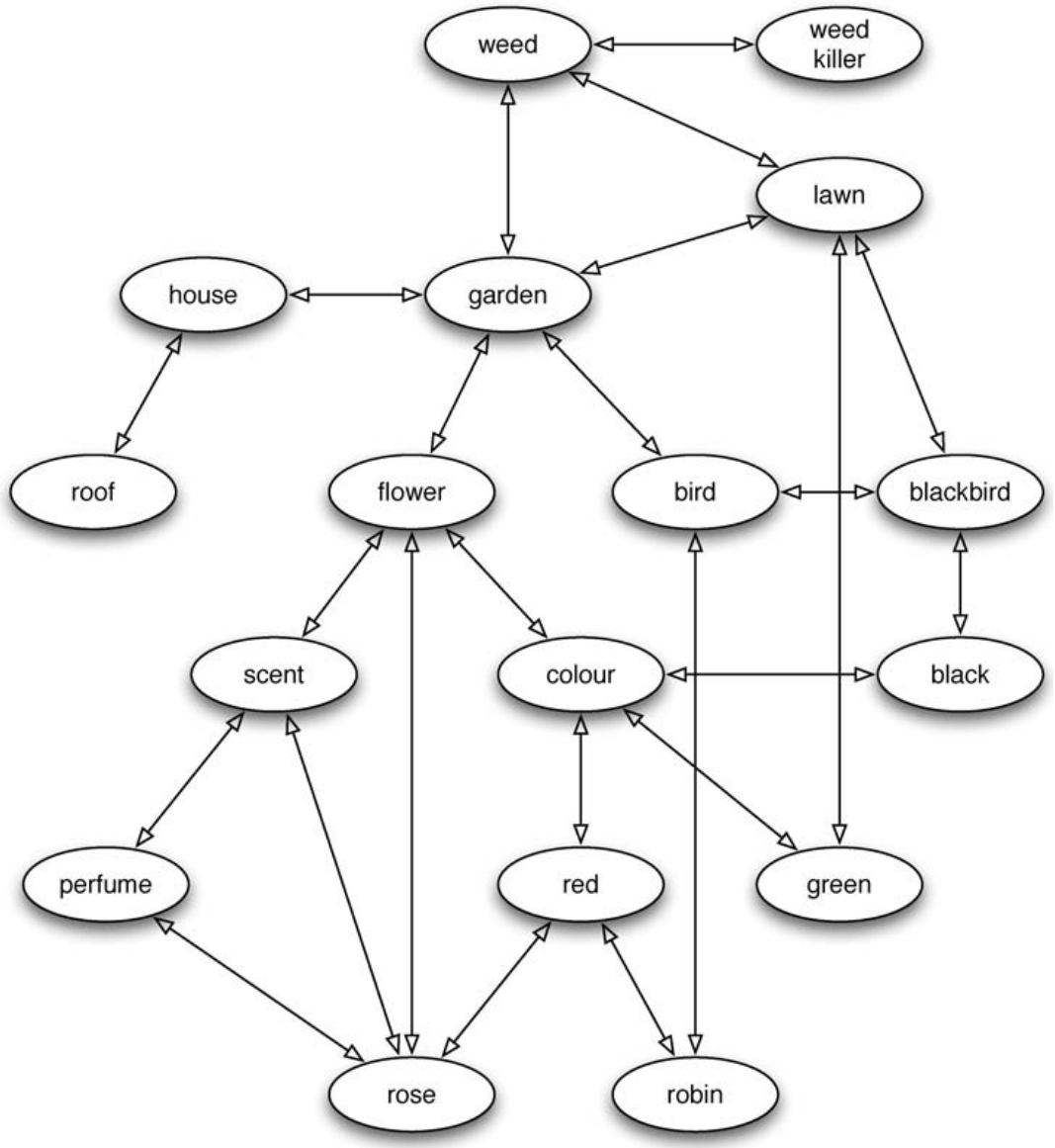


FIGURE 5.2 A small portion of my semantic network – things very quickly get tangled up

"has wings", "flies", "has liver" is stored at each node. Meaning is broken down into smaller units of meaning.

We can take this idea that meaning is best represented by combinations of smaller units of meaning much further so that it acts as the basis for a theory of semantic memory. We call such approaches decompositional theories. It works very well in some domains. The classic linguistic example is that of *kinship terms*. Think about how we might define your relatives relative to yourself. There's some redundancy in the meanings: if you know the gender of your ancestor, the generation, and whether it's maternal or paternal, you can work out who we're talking about. So female, maternal, second – that would be my grandmother Lillian on my mother's side. Male, paternal, second – that would be my grandfather Walter on my father's side. We can represent me, my two parents, and four grandparents – seven people – in terms of combinations of three things. That's parsimonious. We could extend the scheme to include siblings and take the ancestry further back just by extending the things, or *features*, as they're called, that we combine.

The idea of a semantic feature is a useful and powerful one: we represent the meanings of words by different combinations of a much smaller number of semantic features. The scheme has two big advantages. First, for most words, we no longer have to worry about meaning; all we have to worry about now is the meaning of a much smaller set of semantic features. The second advantage is that it's economical. Although economy might not, especially in the light of any conflicting evidence, be essential, it is elegant. It also reduces the circularity problem with our mental dictionary, if not obliterating it. Because of these advantages, feature-based approaches have been popular in artificial intelligence approaches to language and translation. Yorick Wilks (1976) described a computer simulation where the meanings of 600 words were captured by combinations of 80 features. An example, simplified and translated from the computer program LISP into English: the meaning of the verb "drink" is decomposed into "an action done by animate things to liquids causing the liquid to be in the animate thing via an opening in the animate thing". Let's not get too hung up on whether this is plausible; and remember this is a translation from features into a computer language and back to English. The important point is that we can capture the meaning of complex notions in terms of combinations of a small number of simple ones – and it works.

Some linguists argue that the meanings of words in all languages can be described in terms of combinations of a small – perhaps as small as 60 – number of features universal to all languages. The leading exponent of this approach is Anna Wierzbicka (e.g. 1996, 2004), whose major insight is that the complexity of meaning in different languages can be reduced to primitives that are common to all languages. The features include translated items corresponding to I, YOU, SOMEONE, SAY, TRUE, HAPPEN, MOVE, LIVE, DIE, and NOW. Now this is not to say that this way of describing language is how humans actually compute meaning in

everyday life, but it does demonstrate the viability of the approach, and shows how it is possible to extend meaning from beyond the confines of an individual language. Their universality does give an inkling that these sorts of features might be the atoms of thought.

The other great advantage of semantic features is that they give us a way to build up the meaning of sentences. The meaning of a sentence is no longer the meaning of the individual words, but of a combination of semantic features. The combinatorial approach gives us a method of coping with ambiguity and explains why certain combinations of words strike us as anomalous (Katz & Fodor, 1963). The word "ball" is ambiguous. Here it is in three utterances.

> Felix picked up the ball in his mouth and ran towards goal.
> The pet owners' annual ball was held in the field this year.
> The house kicked the ball.

The sense of "ball" meaning "small round object" is the only one that fits with an animate thing picking it up; the "dance" sense only fits with being at a location; and inanimate things can't carry out actions in either sense, so we find the final sentence anomalous.

How do semantic features fare at predicting performance in the sentence verification task? Very well, with some modifications. Rips et al. (1973) divided semantic features into two sorts, *defining* and *characteristic*. Defining features are those that are an essential part of the word's definition. A bird lays eggs and has wings. Characteristic features are usually true but aren't always: a bird usually flies, but it's not an essential, defining feature of birds. What happens when we have to verify a sentence such as "A penguin is a bird"? In Rips et al.'s account, when doing the sentence verification task, we first compare the overall featural similarity of the two key words in the sentence ("penguin" and "bird"). If there's a very high overlap (as there would be with "robin" and "bird"), we should respond TRUE, and if there's a very low overlap (as between "aardvark" and "bird), we should respond FALSE (assuming of course the person knows what an aardvark is). There are, however, pairs of words with a moderate amount of overlap, such as "penguin" and "bird", and "pine" and "flower"; in these cases we are forced to go to a second stage of comparison, where we carefully check the defining features alone.

Feature theories bear much similarity to network models, and the connectionist models we're just about to come to combine the two concepts. It's difficult and perhaps impossible to distinguish between the two, and now certainly not worth the effort. Feature-based accounts of meaning do face a number of problems, however, the foremost of which is that it's straightforward enough to list features for dogs and aardvarks, but what about our old acquaintances truth and justice? Furthermore, some words don't seem to have any defining features; the most

famous example is "game" (Wittgenstein, 1953). What do all games have in common? For everything we can think of, there's a counter-example. Involves opponents? Solitaire. It's enjoyable? Tell that to a footballer or chess grandmaster on a bad day. One response to this difficulty is to ditch the whole idea of defining features, and say that category membership is defined by family resemblance, so that members of categories such as "game" merely resemble each other. But this apparent problem isn't really a problem at all for a feature-based theory; we merely dispense with the notion of necessary defining features, and look at the total amount of overlap.

We're still faced with the problem that we don't really know what the set of our semantic features is, and how we combine them to form the meanings of truth and justice. But I don't think we should worry about this too much; we know from the linguistic work of Wierzbicka that it can be done in principle. We need to move away from the idea that human semantic features have nice, straightforward linguistic correspondences like "has wings", "big", "has liver"; there's no reason at all why they should, and many good reasons why many of them won't. There's also no reason to suppose that a semantic feature is simply either "on" or "of"; it might have a value between one and zero, say. These are all characteristics of the connectionist models I describe in the coming sections, but it is worth reiterating the point that we don't know what our features are, and they might not be easy to express in words. We don't know what's in our heads. A related point is that features such as the sort we're thinking of here provide us with a means out of the terrible self-referential loop that confronts dictionaries because some of our semantic features can be linked to our perceptual systems. We can envisage the mind as a huge network. Words will link to other words and features, back perhaps through many levels of connections, to the sorts of representation with which vision, sound, touch, smell, and taste connect. Not all connections will need to go this far; some might even just connect to other words.

I don't want to give the impression that a theory of meaning is now in the bag. It should be made very clear that not everyone even agrees that decompositional semantics, where we break the meaning of a word down into smaller units of meaning, provides the best account of meaning: the leading alternative view is non-decompositional semantics, which maintains that for every word we know there's a concept that stands in a one-to-one relationship with the word. It's extremely difficult to distinguish the decompositional and non-decompositional approaches experimentally, and the impetus in current research is definitely with the decompositional approach. It's also difficult to see how the non-decompositional approach can be related to perception so readily.

This description sounds vague and speculative, but connectionist models show how it can all work in practice.

There's a rare (for Britain) bird somewhere in this photograph. Or perhaps it's an aardvark

What does neuropsychology tell us about meaning?

The brain is a delicate organ protected by a thick casing, the skull. Nevertheless there are numerous horrible ways in which it can be damaged, of which the most common are head injuries, particularly including missile wounds and brain damage from some car crashes, and strokes, when the blood supply is cut off to part of the brain – brain cells are very sensitive to oxygen starvation and die very quickly without a supply of oxygenated blood. I'll talk more about the range of disasters that can befall the brain in a later chapter, but here I want to focus on a disorder known as *deep dyslexia*. Deep dyslexia is a profound problem with reading that previously competent adult readers acquire as a result of severe damage to parts of the left hemisphere of the brain (Marshall & Newcombe, 1966, 1973). It's characterised by a number of symptoms, including great difficulty in reading aloud pronounceable non-words (often called pseudowords), such as DAT, NITE, SMOUTH, and GRAT. Normally, as I'll show in a later chapter, people have no difficulty in pronouncing strings of letters such as these, and what's more people agree on how they should be pronounced. People with deep dyslexia also find nouns easier to read than adjectives, and adjectives easier to read than verbs. They have particular

difficulty in reading grammatical words, such as "of", "in", "their", "where", and "some", even though these words are very common and usually short. They also make errors that seem to be based on the visual appearance of the word, mispronouncing it for another that looks quite similar, such as saying "perfume" for "perform" and "signal" when asked to read "single", and derivational errors, where they misread the word as another one grammatically derived from it, such as saying "performance" instead of "performing" and "entertain" instead of "entertainment". Perhaps unsurprisingly, they also make what are called mixed errors, where the word said is related to the target in both meaning and sound (e.g. saying "late" for "last").

In these respects deep dyslexia is very similar to another acquired disorder known as phonological dyslexia, but deep dyslexia is characterised by the presence of a very curious sort of error known as *semantic paralexia*. When a person makes a semantic paralexia, they pronounce a word as though it's related in meaning to the one they're trying to read. A few examples should make this clear.

Daughter is read as "sister".
Kill is read as "hate".
Rose is read as "flower".
Sergeant is read as "soldier".

They find words referring to more imageable concepts easier to read than words referring to less imageable ones. Imageability is simply how easy it is to form a mental image of the thing the word refers to. Close your eyes, sit back, and try to form images of these words: "poppy", "cloud", "dog"; now try the same with "justice", "truth", "knowledge". The difference in the ease with which you can form an image should be striking.

Semantic paralexias seem mysterious enough by themselves, but the really interesting thing about these symptoms is that they always occur together in deep dyslexia. If a patient has the defining characteristic of making semantic paralexias, they will also always have difficulty with non-words, grammatical words, and also make visual errors, and so on (although the proportion of types of error might vary from patient to patient). Why should this be? At first sight semantic paralexias look like a completely different sort of thing altogether from visual errors and the type of part of speech. There's no apparent reason why they should all occur together. For some time researchers struggled to find an explanation for this pattern. Deep dyslexia is one of the most severe of the acquired reading disorders, and some researchers argued that it's the outcome of a highly damaged system trying to read normally (Morton & Patterson, 1980). Although this approach sounds very plausible, it lacks detail, and doesn't explain why things go together as they do. Other researchers adopted a completely different approach, noting that as deep dyslexia is found when a patient has extensive damage to the left hemisphere with the loss of

much of the brain that normally deals with reading, perhaps it doesn't reflect the normal reading system struggling bravely on at all, but reflects the performance of a much inferior system. One suggestion was that it reflects the performance of a right-hemisphere reading system that's normally suppressed by the much more effective reading system, but can come to the fore when the usual left-hemisphere system is virtually obliterated (Coltheart, 1980; Zaidel & Peters, 1981). This hypothesis does have some appealing aspects, because we know from various sources that the right hemisphere of the brain can read in a very limited way but makes many semantic errors; however, it can't account for the precise range of symptoms found in deep dyslexia, and in particular cannot explain why all the other symptoms occur in addition to semantic paralexias.

We're not much further forward in explaining the pattern of reading difficulties in deep dyslexia. It took a sophisticated computer simulation of meaning and how we access meaning in reading to illuminate the problem, and also to show how powerful a system based on decompositional semantics can be at explaining normal performance but also how brain damage can disrupt the system. Warning! The next section is hard going, and might need to be read a few times. It's worth it though.

Hinton and Shallice (1991) produced a computer simulation of aspects of reading, focusing on how we get to the meaning of a word from print. In their connectionist model, they used many simple processing units arranged in layers, with each unit in one layer connected to every unit in the level above. In the lowest level there were 28 units representing *graphemes*, the smallest unit of printed language that can make a difference to the meaning of a word. This is more or less a fancy way of saying a letter – such as "c", "o", "a", and "t". Hinton and Shallice used 28 graphemes rather than 26 because they also wanted to represent information about the place of the letter in the word. So the input level is a way of representing the visual appearance of the printed word.

The output level was 68 units, each corresponding to a semantic feature. These features were things like "mammal", "has-legs", "brown", "main-shape-2D", "soft", and "fierce". Now no one would pretend that human semantic features are anything like these, but the underlying principle is the same. These features were sufficient to encode the meanings of 40 short words such as "cat", "cot", "cow", "rat", and "bed". So, for example, the meaning of "cat" would correspond to the positive activation of the semantic features "max-size-foot-to-two-yards", "main-shape-3D", "has-legs", "moves", "mammal", "fierce", "carnivore", among others. The idea is that when we read, we activate the appropriate letters, or graphemic input units, and out comes the right meaning, or pattern of activation on the semantic units.

The model was trained to produce approximately the right output to any given input. In this sort of model the outputs don't have to be exactly right, they just have to be good enough – as long as the output is closer to the target output than

the semantic representation of another word, that's the meaning that will be "accessed". The model was trained using a learning algorithm known as back-propagation. We've met this technique before. Each connection in the network has a weight or connection strength associated with it. We apply activation to the input unit and activation flows along the connections to the output units. Suppose we had a very simple network where we wired whole words directly to semantic features. So when we pressed the switch for CAT the connection strength to "has-legs", "moves", "mammal", and so on, would be +1, ensuring that those semantic features then light up at the other end, and the connection strengths to "sweet", "tastes-strong", and "made-of-wood" would be zero, because we don't want those features to light up. The Hinton and Shallice network is much more complicated than that, because it has to learn to produce the right semantic outputs for 40 words based on the input of their letters. So the input pattern CAT has to activate a very different output pattern from COT even though there's a lot of overlap in the letters. In addition, to learn material like this it's been shown that we must have an intermediate level of units (called the hidden units) that just mediate between the input and output. Hinton and Shallice used 40 of these (the exact number isn't critical). That means that the complete network had nearly 4000 connections in it – which is one reason why this sort of approach wasn't really feasible until powerful computers were readily available. The network starts off with random connection strengths (between plus and minus 0.3), and would therefore at first produce utter rubbish. It's trained to produce the right patterns, more or less, by repetitively presenting all the inputs to the network, seeing what the network outputs for that particular input, comparing it to what it should output, and gradually changing the connection strengths so that the next time the output it produces is a bit more like what it should be. It then moves on to the next input. This whole process is then repeated perhaps many thousands of times until a criterion of good performance is reached: the network produces the appropriate semantic output for each word it's presented with. The network has been taught to read.

No one is claiming that humans learn anything by back-propagation, although the process of learning by gradually reducing the errors we make has some plausibility to it. The important thing is that back-propagation provides modellers with a means of constructing networks without having to craft every aspect by hand. Indeed, with a network of this complexity, it would be near impossible to do so.

In fact the model was even more complicated. There's something unsatisfactory about being so dictatorial about exactly what the semantic units should be doing. It would be nice if the system could learn something about how the semantic units are related to each other – for example, that mammals can move but can't be made of glass. We want the semantic units to have some interdependencies. Also, at first the network didn't work terribly well with the structure I've just described. It tended to confuse similar inputs – it tended to create semantic outputs for words

like "cat" and "cot" that are visually similar, unless the network was given a huge amount of learning. Hinton and Shallice therefore introduced another layer of units, connected just to the semantic features, called "clean-up units". In fact there's a feedback loop between the semantic units and the clean-up units (Figure 5.3). This type of network is called a *recurrent network*, and the result is a system that can

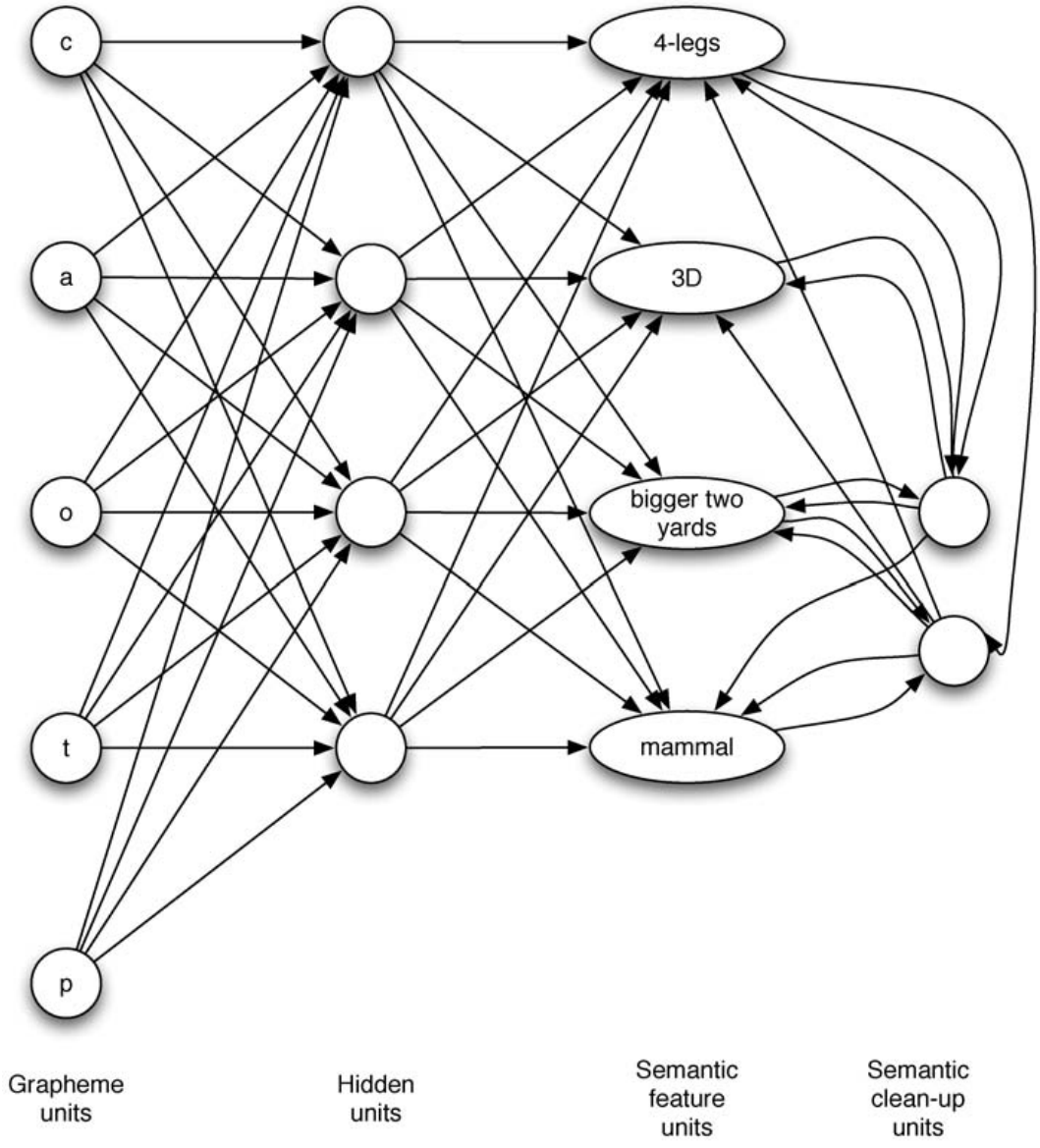


FIGURE 5.3 Simplified portion of the Hinton and Shallice connectionist network model of deep dyslexia. The figure gives some idea of the complexity of the model. There's an input level of graphemes representing the visual appearance of the word, a level of hidden units essential for the model to be able to learn, and an output level of semantic features representing a word's meaning. The semantic features are connected to clean-up units in both directions, enabling the semantic system to develop structure with experience

learn more efficiently and develop its own semantic structure, learning regularities in the semantic representation of words.

Now we have a system that has both learned to read and developed its own semantic representations. It's a functioning adult reader, albeit in a very limited way. How can we simulate the effects of brain damage? There are several ways, which we can call *lesioning* the network – in just the same way as brains are lesioned. We can randomly reset some of the connection weights to zero, or a random value, and of course there are types of connection (input to hidden, hidden to output, and semantic to clean-up) we could disrupt. The details of how we lesion turn out not to be too important, although damage to the connections involving the clean-up units is the most interesting.

The advantage of connectionist models such as Hinton and Shallice's is that they display what we call *graceful degradation*. If you remove one transistor from your computer, it will stop working altogether. That's not graceful. But if you take a trained connectionist network and damage it just a little, tiny bit, you probably won't notice any difference. As you increase the amount of damage, the network starts to make errors. What's more, these errors aren't random, as we shall see. The more you damage it, the more errors it makes. Of course, there comes a point when it behaves terribly, and another when it behaves randomly; if you zero every connection then it won't work at all. But this degradation of performance is a gradual one.

When you inflict a moderate amount of damage to the Hinton and Shallice network, it makes errors like the semantic paralexias of a deep dyslexic. For example, given the visual input corresponding to the word "cot", it produces the semantic output corresponding to (or at least closest to) the word "bed"; it might say "dog" for "cat" and "hip" for "rib".

Perhaps this outcome isn't too surprising; after all, the network was trained to associate meaning with visual appearance, so damage to it is going to disrupt that matching. It's the fact that the lesioned network systematically produces semantic errors, rather than some random output, that's intriguing. But that is by no means all: it also produces visual errors, such as saying "cot" for "cat" and "log" for "dog". Now that is surprising – why should the visual appearance of words matter?

The explanation illuminates the way in which humans represent meaning. Remember that the clean-up units allow the semantic features to discover structure among themselves. It's this structure that's important. We've moved away from the simple presence or absence of semantic features to something much more complex and interesting. The structure of our semantic space is best thought of using a visual analogy or two. Imagine you have a bowl and drop a marble into it. It lands on the side. Where will it end up? It will of course roll to the bottom of the bowl. It's obvious that wherever you drop the marble into the bowl, wherever it hits the side, it will roll to the bottom. The bottom of the bowl is called an *attractor*. Semantic space in Hinton and Shallice's simulations ends up being structured in the same way; you can think of semantic space as a landscape, seen from a plane, a rolling

countryside with many valleys separated by mountains. If from your plane you drop a football, it will land and then roll to the lowest point in the valley nearby – the nearest attractor. The lowest point in the valley is the semantic attractor, and the football the input from visual processing. What brain damage does is to change the landscape, at first by eroding the hills and mountains between the semantic valleys. The erosion could lead to the ball ending up in a different place. When you see "cat", the ball should end up in the semantic attractor corresponding to the meaning of "cat", but damage to the scenery means that it might land in the basin of another meaning; remember it's not where the ball lands that's important, but the attractor. It's obvious that valleys nearby in the landscape will be of closely related meanings, so although the ball for "cat" lands in the same place, it might end up in a damaged landscape at the bottom of the valley for "dog" or "rat".

I think this account of why we get semantic errors is straightforward and intuitively clear, but it's far less obvious why we should get visual errors as well. The key point to understanding the explanation is to appreciate the importance of it not being where the ball lands that's important, but where it ends up, and that balls landing in places that are very close together could end up a long way apart. The key insight is that the system learns in a way such that the visually similar representations (like cat and cot) initially point to quite close points in semantic space; the clean-up units then ensure that the ball falls to the proper valley. It's as though the scenery has some wide valleys separated by peaks, and the ball is directed on to the peaks. These are some way away from the valley bottoms, but gravity will do its work and guide the ball to the appropriate valley bottom. But damage to the system erodes the area where the balls first land, so once again, although they land in the same place, they will fall into the wrong valley, but because of the initial mappings this valley might be one of a word related in appearance or one related in meaning. Of course, a word related in both meaning and appearance will have even more potent attractors, explaining why we get so many mixed errors.

The "landscape" is much more complex than this analogy of a landscape suggests, and we would find it impossible to visualise. Because each semantic feature can have an activation level that varies continuously, our semantic landscape has as many dimensions as we have features. It sounds like one of those fantastic theories of modern physics. It does though help to explain what might be the puzzling feature, that landing at an initial point in semantic space can take you to both visually and semantically related attractors. This puzzle is only a problem in two- or three-dimensional space; this behaviour arises as a consequence of the geometry of multidimensional space. (I did say psycholinguistics was hard.)

The model can explain another striking aspect of deep dyslexia: the imageability effect, where words are more likely to read correctly the easier it is to form a mental image of the related concept (so "rose" is more imageable than "abstraction"). In this approach, the high-imageability word has more active semantic features underlying its meaning than a low-imageability one. The more concrete

and imageable a word, the richer is its semantic representation. So in the model the word "post" needs 16 features to specify its meaning, but the word "past" has just two (has-duration and refers-to-time). Semantic representations underlain by many features are going to be more robust to damage, and will be able to pass on more activation to the next processing stage, that of producing the sounds of the words; semantic representations underlain by few active features are less robust and can pass on less activation.

I've spent some time with this model because although it is a model of a specific neuropsychological disorder, it shows how we can develop the semantic feature approach to construct a coherent and plausible account of how the mind deals with meaning. If we scale up the model to human cognition, the semantic representation becomes even more complex, but the same principles apply. The underlying semantic features will be much more numerous and, as I said earlier, likely to be abstract and not necessarily with any straightforward linguistic correspondence. They also provide contact with the perceptual system, which "earths" the semantic system in the real world. Word meanings still correspond to semantic attractors. What happens in language acquisition is that the child abstracts semantic features, and the attractors gradually come to resemble those of adult speakers. Note that in this approach not everyone's semantic space will look exactly the same; all that needs to happen is that our attractors correspond enough for us to be able to communicate. In the terminology introduced earlier, the attractors are the denotations of meaning. Differences in the shape of semantic space give us differences in connotations. This variability is in fact desirable; we all have slightly different associations to particular words, and this approach captures those differences. You might be very fond of tarantulas, although the thought makes me shiver, but we can still talk about them (just) and be confident that we're talking about the same thing.

How can we explain what goes wrong in dementia?

The semantic feature approach explains another neuropsychological disorder in a very straightforward way. The neurodegenerative diseases that fall under the umbrella term "dementia", of which Alzheimer's disease is the best known, display a number of psychological problems. Although the exact cause of dementia isn't known (and there are several types, and there are probably several different causes), the basic pattern is the same: a progressive loss of cognitive and motor functions. The brain of a person with Alzheimer's disease shows a loss in the number of neurons and the presence of tangles and plaques (where the nerve cells become bunched up and knotted together, and surrounded by dead cells and deposits of protein). The earliest symptoms of Alzheimer's disease include subtle difficulties in planning and loss of memory, particularly for recently

learned things. Language is noticeably affected by dementia from quite early on, with difficulty in remembering names and a diminishing vocabulary; the grammatical rules seem relatively well preserved. It's this loss of vocabulary and the increasing inability to remember the names of things that's of interest here. (Don't worry, some difficulty in remembering names is an aspect of normal ageing, too.) Given pictures of common objects to name, people with moderate dementia will struggle. What could explain the difficulty in naming and the shrinking vocabulary? Given that dementia involves progressive loss of neurons, I think the most obvious explanation is that the loss of brain matter means that the person is losing their semantic features.

Researchers have modelled the effects of the progressive loss of semantic features with a connectionist model that shares many feature of the Hinton and Shallice model. Tippett and Farah (1994) constructed a model centred around 32 semantic feature units. These were connected on one side to 16 units that represented the spoken names, and on the other to 16 units that represented the visual appearance of a small set of objects. The model was trained so that activation of an input pattern corresponding to a particular object gave rise to the right pattern of semantic activation, in turn producing the correct name for that object in the name units. Similarly, the model was trained so that the activation of the name units gave rise to the appropriate semantic representation. After training, the model was damaged or *lesioned* by removing semantic units at random.

As you would by now expect, destroying semantic units impaired naming, and the degradation in performance was graceful in that a small amount of damage impaired naming ability, but didn't destroy it completely. As the amount of damage increased, naming performance deteriorated. But the beauty of connectionist models is that they show us how things we might not think are related can in fact be so. The lesioned network had more difficulty with less common names than more frequent ones (frequency had been implemented by giving more training to high-frequency names than to low-frequency ones). The damaged network became very sensitive to the clarity of the visual input pattern; if a weak pattern was presented, corresponding to a degraded image, naming was even worse – and it's been known for some time that the naming ability of people with dementia is sensitive to the quality of the picture; they do better with colour photographs than black-and-white photographs, which in turn lead to better performance than line drawings. Finally, naming could be improved by providing a bit of help with the sound of the name; naming by people with dementia is improved by giving them the hint of the initial sound of the word ("l" for "lion", a technique known as *phonological priming*). So a simple model can give rise to sophisticated and realistic behaviour. The central idea, though, is that the vocabulary and naming problems in dementia are caused by the progressive loss of semantic features.

How is semantic memory organised?

One of the most peculiar neuropsychological deficits of meaning was first studied in detail by Warrington and Shallice in 1984. They noticed that their patient JBR performed much better at naming inanimate objects than animate ones. So he was much better at naming pictures of vehicles than of animals. His difficulties went beyond naming, though, because he also found it more difficult to understand words denoting living things than non-living things, matching the right picture to the name, and even producing a gesture appropriate to the word. This pattern of results, where a patient shows good performance with members of one semantic category and poor performance with another, is called a *semantic category-specific disorder*. It later emerged that other patients show the reverse pattern, performing better with animate things compared to inanimate (Warrington & McCarthy, 1987). The most obvious explanation for these findings is that knowledge of living things is stored in a different part of the brain from knowledge about non-living things.

Although category specificity involving the living–non-living distinction turns out to be a relatively common one (although these disorders are in absolute terms rare), several more specific disorders have been found. One patient had particular difficulty just with fruit and vegetables (Hart et al., 1985); others show problems with proper names (Semenza & Zettin, 1988).

At face value, these deficits suggest that special parts of the brain store different types of information: knowledge about living things in one part and about non-living things in another. It was immediately apparent that this simple explanation was unlikely to be right. JBR's naming was more complex than this initial picture suggests. For a start, he was good at naming parts of the body, even though these are parts of living things. He was also poor at naming several types of non-living things: musical instruments, precious stones, types of material, and foods. The reliable co-occurrence of these categories with living things is difficult to explain: why should damage to the part of the brain storing knowledge of living things also lead to problems with musical instruments and gem stones?

One explanation is that these categories have something in common, and what the brain damage has disrupted is the processing of this shared characteristic, rather than knowledge of the categories themselves. What might this common characteristic be? What do living things, gemstones, foods, musical instruments, and materials have in common that other artefacts don't? One possibility is that we distinguish and describe the first group primarily in terms of their appearance, while we describe artefacts in terms of what we do with them. To give a very simple example, think about how you'd define a "giraffe": it might be something like "an African mammal with a very long neck and long legs and a tan skin with spots and little horns and a happy face that chews the leaves from the top of the woodland canopy", and then contrast that with a definition for a "chair", which might be

something like "a piece of furniture with a seat and back used for sitting in". Contrast "diamond" ("extremely hard, highly reflective form of carbon") with "hammer" ("a hand tool with a head that's used for striking things hard"). I'm not saying that the animate and related things are defined just in terms of their sensory attributes and the artefacts their function, but that animate things depend relatively more on sensory information (Warrington & McCarthy, 1987; Warrington & Shallice, 1984). This observation is supported by an analysis of a large number of dictionary definitions: for living things, the ratio of perceptual to functional attributes is just under 8 to 1, but for non-living things it's much lower, about 1.5 to 1 (Farah & McClelland, 1991).

The idea that different categories depend differentially on sensory and functional information is called the *sensory-functional theory*. So what gets damaged in these patients isn't the categories themselves, but the ability to access the sensory or functional information that underlies them. Imaging studies of blood flow in the brain show that the temporal lobes of the brain don't respond differentially to living and non-living things, but different parts of the brain do respond to perceptual and non-perceptual information (Lee et al., 2002).

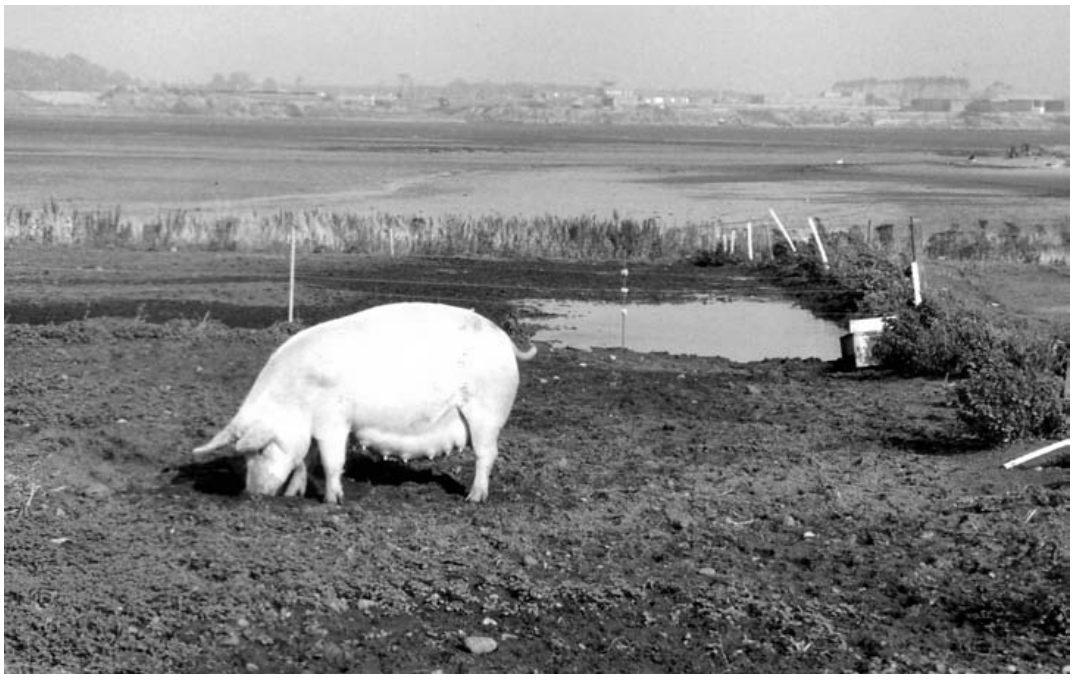
It wouldn't be psycholinguistics without a good argument. The American neuroscientist Alfonso Caramazza has produced robust criticisms of the sensory-functional theory. If the theory is correct, he's argued, a patient who performs badly on living things does so because of a problem in processing sensory features, and therefore will perform badly on all tasks involving living things, and on all other categories (the musical instruments, gemstones, materials, and so on) that also depend heavily on sensory information. But this isn't always the case (Caramazza & Shelton, 1998). There are patients impaired at tasks involving animals but not foodstuffs, whereas others are impaired at foodstuffs but not animals; some patients are impaired at animals but not musical instruments; and knowledge of animals can be spared or damaged independently of plants. It is of course conceivable that some categories rely more on particular sorts of sensory information than others, so what is damaged isn't a wholesale ability to deal with sensory information but just particular types of sensory information, but such an approach would need much more spelling out to be convincing. Caramazza and Shelton also point out that while the idea of sensory information has some coherence (it's just what things look, taste, smell, feel, or sound like), the category of functional information is much less coherent. The concept of "what something is used for" is very restricted; is all non-sensory information functional? What about "a long neck to reach the canopy"? "Lives in the desert"? They also point to imaging studies that suggest that different parts of the brain are in fact differentially activated when processing animals and other categories. Put very broadly, knowledge about animals is stored more towards the back of the lower left temporal lobe of the brain, while knowledge of tools is stored more towards the side, where the temporal, occipital, and parietal lobes of the brain meet (Caramazza & Shelton, 1998;

Vigliocco et al., 2004). An alternative account, the *domain-specific knowledge hypothesis* (thankfully often abbreviated to DSKH), says that because of the obvious evolutionary importance of distinguishing between living and non-living things, the brain has evolved separate mechanisms for dealing with them. So on this account knowledge about different categories is processed in different parts of the brain. Further evidence for some genetic basis to the distinction between living and non-living things comes from the study of a 16-year-old boy known as "Adam", who suffered a stroke the day after he was born. Adam has great difficulty recognising and retrieving information about living things; the fact that the damage occurred so early rules out the possibility that any learned information could have been affected. We are born with different neural systems to store knowledge about living and non-living things.

Clearly we have much to learn about how the brain represents knowledge. I don't find *where* something is stored that useful or interesting or interesting in itself. But neuroscience can tell us a great deal about the principles upon which human knowledge is constructed and stored.

What are statistical models of meaning?

The idea that everything is linked to everything else is deservedly enjoying vogue in the popular science press. Psycholinguists, with their semantic network model of



A bundle of mainly perceptual features

A bundle of mainly functional features

meaning, got there first – or at least early on. I've shown how connectionist modelling provides an account of meaning at two levels. At the the lower level we have semantic features, which capture aspects of meaning, and which are the atoms or "primitive" of thought. These are interconnected in an enormous mental network, with connections to words in one direction and sensory representations in the other. The features are also connected to each other in a way that enables the mind to discover structure and regularity among them. This ability enables us to view semantic representations at a higher level as multidimensional landscapes, with peaks and valleys, and the meaning of words corresponding to semantic attractors.

An approach that's very similar in spirit is called *latent semantic analysis* (Landauer & Dumais, 1997). The motivation of this approach is that meaning arises from co-occurrence. As we're exposed to language, from infancy on, some words tend to occur frequently with other words, some occasionally with other words, and some combinations occur rarely or never. For example, "doctor" and "nurse", "bread" and "butter", and "dog" and "cat" are highly associated; "proton" and "xylophone" might never have occurred in proximity before I wrote this (although rather to my amazement a Google search with both terms present in

pages generated over 5,300 hits). For any word, we can identify the frequency of occurrence of all others within a certain distance.

Landauer and Dumais worked out (using a computer; this approach is another that would be unthinkable without a good fast computer) how often every word in an encyclopaedia co-occurred with every other word in the same encyclopaedia entry. The encyclopaedia contained over 4.5 million words, made up of 60,768 different words, arranged in 30,473 entries. You could represent this co-occurrence in a huge, 60,768 by 60,768 grid. Each cell would signify how many times that word pair co-occurred in the same encyclopaedia entry (the score for xylophone and proton probably being zero). They then simplified this down to 300 dimensions. Words with similar patterns on these 300 dimensions should have very similar meanings, and highly similar patterns should mean that the words have very similar meanings indeed – that is, they're synonyms. Unfortunately "xylophone" doesn't have synonyms, but you can play with online synonym generators to discover that synonyms of "happy" include "blessed" and "blissful", and those of "miserable" include "abject" and "wretched". The synonyms generated by their model corresponded very well with published lists of synonyms. We know that we're faster to recognise a word if it's preceded by one similar in meaning – so we're faster to identify "butter" if we see "bread" immediately beforehand. The degree to which recognition is speeded up depends on how closely related the two words are, and this distance, and therefore the magnitude of the recognition benefit, is predicted quite well by this sort of multidimensional analysis of 300 million words taken from online messages (Lund et al., 1995).

This approach captures aspects of meaning, but it's a much bigger step to say that this is how meaning originates in humans. Something seems to be missing from the account: the real world. You can't learn language just by listening to a stream of words; children pay a great deal of attention to the environment – as indeed we all do. Meaning is more than association to other words; it's grounding in perception and action too. This claim does not deny that the shape of semantic space isn't modified by co-occurrence information – it's probably one of the sources of information that influences the way clean-up units structure semantic space. This type of statistical analysis is interesting and useful, and says a lot about how words connect to words, but not how words connect to the world.

What's grounding?

Even the best dictionaries have an element of circularity about them. However careful the lexicographers are, eventually they have to end up defining words in terms of other words. The same limitation is present in computer databases. And it would be very difficult to take a robot seriously unless it had some way of taking in information from the environment in real time, such as robo-eyes and robo-ears, no

matter how wonderfully structured and enormous its database. Humans are special in that meanings ultimately connect with the world. I've said several times before that our internal representations are *grounded* in our perceptions, actions, and feelings. Concepts have very direct links to the world (Barsalou, 2003, 2008; Glenberg, 2007). Our minds don't work in isolation – they are *situated* within the world. According to this view, concepts and meaning aren't just abstract, amodal things: thinking about real-world objects, for example, involves the visual perceptual system. Furthermore, according to the situated cognition idea, concepts are less stable than has usually been thought, varying depending on the context and situation. Barsalou (2003) had people perform two tasks simultaneously: using their hands to imagine performing some manual operations, and identifying the properties of concepts. Sometimes the actions being performed were relevant to the concepts being described, in which case the participants were more likely to mention related aspects of the concepts. For example, if they were performing the action of opening a drawer, they were more likely to mention clothes likely to be found inside a clothes dresser than otherwise.

There is some evidence that our mental situation in the world takes a very concrete form, in that there are direct links between representations of perceptions and actions. What happens in the brain when we hear the word "kick"? We see Wernicke's region, the part of the left temporal lobe of the brain that we know plays a vital role in accessing word meanings, light up like a Christmas tree when seen using brain imaging. It would be worrying if it didn't. We also see some activation in Broca's area, a region towards the front of the left hemisphere that we know to be involved in producing speech. Perhaps there are echoes here of the behaviourist idea that thought is language, but this result isn't too surprising. What is very surprising is that the functional magnetic resonance imaging (fMRI) scans show that there is activation in the parts of the brain that deal with motor control – and the motor control of the leg at that (Glenberg, 2007; Hauk et al., 2004). It's as though when we hear "kick", we give a little mental kick. Similarly, if we hear a word like "catch", we see some activation in the parts of the brain that control the movements of the hand, and if you hear "I eat an apple", you get activation of the parts that control the mouth (Tettamanti et al., 2005). This motor activity peaks very quickly – within 20 milliseconds of the peak activation in the parts of the brain traditionally thought to be involved in recognising words and processing meaning (Pulvermüller et al., 2003), which is so fast that it would appear to rule out the explanation that people are just consciously reflecting on or rehearsing what they've just heard. This idea that thinking or understanding language causes activation in the parts of the brain to do with how the body deals with these concepts is called *embodiment*. Language is grounded to the world, and that grounding happens in the parts of the brain that deal with perception and action. Just like you'd think.