In R. Morelli, W. Miller Brown, D. Anselmi, K. Haberlandt & D. Lloyd (1992). Minds, Brains & Computers: Perspectives in Cognitive Science and Artificial Intelligence, Ablex Publishing Corporation: Norwood, NJ.

CHAPTER SEVEN

Can Connectionist Models Discover the Structure of Natural Language?

James L. McClelland

Department of Psychology
Carnegie-Mellon University
Pittsburgh, PA

INTRODUCTION

When it comes to building a cognitive system, where should we start? Should we start with symbols and rules, or should we start perhaps with something different? The approach that our research group has been taking is to start with units and connections, and to have the symbolic activity of the system be an emergent property of its behavior rather than something that is built in in advance by stipulation.

Symbol-processing theorists (e.g., Fodor & Pylyshyn, 1988) have argued that either the Connectionist approach is unable to capture the structure of cognition, particularly the meaning of sentences and other kinds of thoughts, or Connectionist models are mere implementations of the symbols that are the centerpiece of the classical symbol-processing architecture. I would like to point out some of the shortcomings in the architecture that the symbol-processing theory proposes as the substrate for cognition. I will give you a brief statement of the classical symbol-processing view of cognitive architecture, focusing on what are its essential characteristics, and I will point out that there are aspects of human cognition that are not really dreamt of in the symbol processing framework. Then

I will characterize an alternative architecture, specifically those features which a cognitive architecture should have, which, I believe, are lacking in the symbol-processing framework. Next, I will tell you about a model developed in my laboratory that we hope will begin to capture these kinds of characteristics. I will also show you how it captures some of the aspects of natural language which have often been set aside, or considered less important because they couldn't be dealt with until now.

THE COGNITIVE ARCHITECTURE OF THE SYMBOL-PROCESSING THEORISTS

According to symbol processing theorists (e.g., Fodor & Pylyshyn, 1988), the cognitive architecture should have a combinatorial syntax and semantics. That means the following: First, a combinatorial representation should be either an atom or a molecule. Second, a molecule itself consists of a bunch of constituents each of which itself may be an atom or a molecule; at the bottom there are only atoms. Atoms are basic primitive symbols, and the molecules are just combinations of atoms arranged in a structured way. Third, the content of a molecular representation is a function of the semantics of its parts—that is, the semantics of the molecule is a function of the semantics of the parts of the molecule and of the organization of the constituents.

The processes that interpret these things and assign them their meanings apply to representations by reference to their form or structure. For example, a rule that applies to a molecular representation (some sort of structural description of something) refers to its structure and produces a resulting interpreted output based on that reference to the structure. According to symbol-processing theorists, these principles underlie both language and thought. Moreover, they take linguistic capacity to be a paradigmatic case of systematic cognition.

My own research has been closer to the area of language than to the area of general problem solving. Therefore, I will restrict what I have to say to a discussion of the use and the extent of the applicability of these principles to language rather than to things like reasoning. I am going to argue that these principles really don't work. There is a lot of meaning that is missed by the notion that the semantic content can be captured in this architecture.

The symbol-processing theory argues that there are three characteristics of the cognitive processes that human beings engage in that motivate this approach. One of these characteristics, which will be the main focus of my chapter, is called compositionality. *Compositionality* states that a word makes approximately the same semantic contribution to the meaning of every expression in which it occurs. If you have a sentence that has a particular word in it, the contribution the word makes to the meaning of that expression is the same across all the different expressions in which that word can occur. This semantic contribution

is what is called the word's meaning or semantic value. The notion is that each word has a meaning, and that meaning shows up in the meaning of each of the sentences of which that word is a part. These meanings are the constituents of the meanings of the expressions. That is what allows the rules of language to refer to the forms of expressions that construct meanings, because the meanings of the individual parts merely have to be arranged properly in order to get things to work. I do not believe that compositionality correctly describes the way people understand sentences. Maybe it is acceptable as a rough, crude, first approximation, but I will try to argue that it is really missing a lot.

Problems with the Symbol-Processing Theory

There are two problems with the above view. First of all, there are representations that conventional architectures have to treat as atomic because they can't be strictly analyzed. Nevertheless, they do have structure. Second, the semantic content of an expression can only be partly characterized by rules that apply to representations in respect to their form. Let me begin with a very simple example dealing with the topic of derivational morphology, that is to say, of how the meanings of words get formed out of parts of words. We will move from parts of words to whole words later on and talk about how words contribute to the meanings of sentences.

Consider the words shown in Figure 7.1 that all start with the prefix *pre-*. Certain of these words, like *preview* and *prefabricate*, can reasonably be argued to consist of two meaning components one that means before and comes from the prefix *pre-*, and one that means whatever the rest of the word means. So, *preview* means to view before, which seems perfectly reasonable. *Prefabricate* means to fabricate before. That's also fine. At the other end of this list, we have words like *prefer* and *pretend*. Perhaps we can see some role for the *pre-* in *prefer*, but it seems more reasonable in these cases to say, well, these are just atoms. Especially in *pretend*, neither of the two parts by themselves are really contributing to the meaning of the whole. It is just an atom.

But what about the words in the middle, *predict* and *presume?* Well, I think there is pretty good reason to think that they occupy an intermediate position in this respect. *Predict* has a sense of beforeness, but at least *dict* doesn't really have what we might take to be something like its conventional meaning in this compound. So, it doesn't just mean "to say before," as we might think, if we just assign some meaning to *dict*. It means "to anticipate the future," "to say what is going to happen later," so it has more to it than just saying before. We don't use *predict* if I just happen to utter a word right before somebody else says the same word. So, there is something about the meaning of *predict* that isn't predictable directly from its component parts!

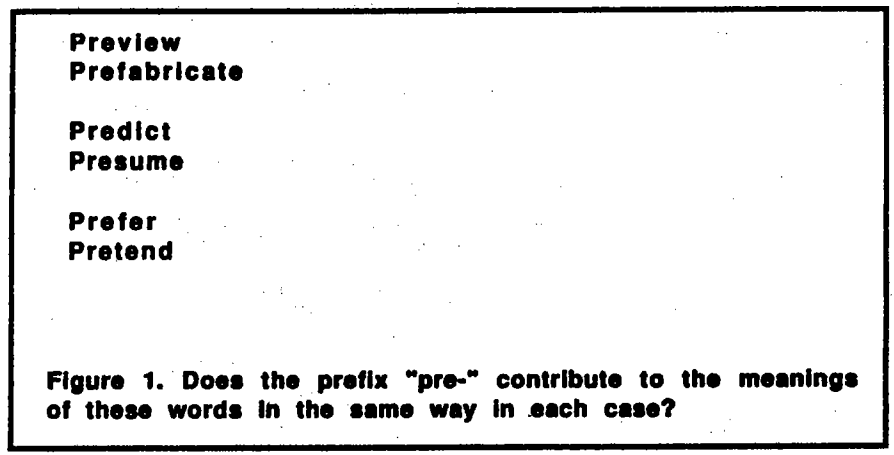

Figure 1. Does the prefix "pre-" contribute to the meanings of these words in the same way in each case?

Figure 7.1. Does the prefix *pre-* contribute to the meanings of these words in the same way in each case?

On the other hand, there is something about the meaning of *predict* that is related to the meanings of its parts, and this is the dilemma that I think the symbol-processing view is faced with. If *predict* is treated as an atom, we lose the fact that is has internal structure that gives some indication of its meaning. On the other hand, if it is a molecule, we lose the fact that its meaning cannot be completely predicted from the meanings of its parts. So the notion that we either have atoms or molecules doesn't quite work in this case, and the notion that we could make up meanings of wholes by using (syntactic) rules to compose the meanings of the wholes isn't quite right. It only covers some of the cases.

What is really happening here? This is where I will begin to give you a flavor for an alternative approach. What happens, it seems to me, is something roughly like the following: Words get coined when things come along for which there are no existing words. Speakers find ways of giving clues to what they mean by combining morphemes. They're hinting to the listener. Morphemes don't have meanings; they are clues to the meanings of the whole word. Listeners use these clues, together with what they know about what the speaker is saying, to figure out what the speaker means. There are other clues that go together in the process with the clues provided by the parts of the words. The parts are, therefore, clues to the meaning of the whole. They are not constituents of the meaning of the whole.

We have talked about the parts of words and how they go together in influencing the meanings of words. But the symbol-processing theorists are more concerned with sentences. It has often been argued that, once a word has been coined, it becomes frozen in the lexicon. They might concede that the "cue"

notion I have described has something to do with how the word is formed in the first place. But after that, it is just in the lexicon; you memorize it and it acts as an atom. There is important work on morphology that has this character, but this will not work in sentences, because we all know from Chomsky's (1959) critique of Skinner that we cannot expect people to have heard sentences before, and yet people can understand sentences. How can that be? The classical approach says it's by virtue of compositionality: You compose the meaning of the whole out of the meanings of the parts together with the rules.

Take the following sentences:

(1) John loves the girl.
(2) The girl loves John.

The classical theorist notes that the word *loves* means the same thing in each of these two sentences; that is, the relationship that John bears to the girl in the first sentence is the same relationship that the girl bears to John in the second sentence. But what about these sentences:

(3) The Pope loves sinners.

Is it the case that the Pope bears the same relationship to the sinners in this sentence as John bears to the girl in the first sentence? Quite possibly not. We hope not, but who knows? I mean, the Pope is allowed to have regular emotions, too; or maybe John is a Pope. Who knows? It is certainly not necessarily the case. Now consider yet another example.

(4) The Pope loves ice cream.

Now probably the Pope's relation to ice cream is different from his relation to sinners, and from John's relation to the girl. It's not the same relationship. If you look in the dictionary under the word *love*, you will find that it has many, many different senses listed. In fact, it's quite clear that words, verbs in particular, are widely polysemous. The different usages tend to be related by a family resemblance, but they are not all the same. They are not just merely different spellings for unrelated concepts though, like the word *ball* as in round spherical object and the word *ball* for a fancy dance. For the word *love*, these are related concepts and we use the same word because it is a good clue to the class of concepts of which we are thinking. But the particular concept that is implicated in any particular utterance is dependent upon the rest of the sentence. In these sentences we can see that the word *loves* sometimes depends for its actual meaning on the object of the sentence. John and the Pope probably have the same relationship to ice cream, but you can see that the meaning of *loves* could well depend in part on the object—viz., ice cream.

We might say, then, that the classical view has overstated the case a bit. Perhaps we have to make the rules sensitive to content as well as form. Many people have tried to save the classical notion of composing meanings by writing context sensitive rules. Here's one: *Loves*, for example, might be said to acquire the features ["+ compassionate"] and ["- erotic"] in the context of a religious subject and an object which is marked as a sinner. This would be a context-sensitive rule that would refer both to structure and to content in formulating the meaning that the word takes in the sentence as a whole. Now, this may be a slightly irreverent example, but I would like to suggest that such a rule as this doesn't work very well for the sentence:

(5) Jimmy Swaggert loves prostitutes.

As you see, you can have a religious subject and a sinful object without necessarily having nonerotic and compassionate love!

We see then that the meaning of each part can depend on the rest of the whole sentence in which the part is embedded. In fact, what you happen to know in terms of your background knowledge comes into play in the way you interpret things. We happen to know something about the people in the previous sentence, and that influences our reading of the sentence.

I am now going to go briefly through several phenomena that compositionality seems to miss.

Polysemy. We have already talked about the first, which has to do with the polysemy of a word. Remember that the symbol-processing theorists have the option of saying, "Well, each of these is just a totally unrelated atom, so it is a mere problem of figuring out which of these unrelated atoms is the one we have to insert in each of these particular cases." In this case they forget that those meanings are all related to each other. Or they have the other option of ignoring the fact that there really are differences in the meaning of the cases. Let me show you several other examples to illustrate how general and ubiquitous these problems are.

Shading. Words get specified by the context in which they occur. The word *container* is a fairly generic specification of some kind of an object. When we put it in context such as in (6):

(6) The apples were in the container.

We impose certain further constraints on this container. Likely enough, it's not too small. It may well be porous. In fact it's sort of useful if it is, because then apples can breathe and they don't get moldy so fast. Whereas if I use it in a sentence like

(7) The cola was in the container.

it would be a poor soft-drink container that had the same properties as an apple container. Psychological experiments have been done to show that, when people read such sentences, they store in memory a more specified version. When it comes time to retrieve this later, if I cue you with *basket*, it's a better cue to (6) than the word *container* is. The context further specified the meaning here, and that's what got stored away in memory.

(8) The baby rolled the ball to her daddy.

Consider the above sentence. We have this nice word *ball*. Obviously we are not talking about a fancy dance here, but what I'd like to think in this case, and I hope you share this intuition, is that the kind of ball that we envision when we think about the event described by this sentence is rather different from the kind of ball we think about when we imagine Jose Conseco pounding the ball into the bleachers. The context here shades the characteristics this ball is likely to have, changing its features, making it perhaps slightly larger and squishier and more likely to be multicolored in this case than in the Conseco example.

(9) The girl played the piano.
(10) The man lifted the piano.

Emphasis. When people hear a sentence such as (9), the word *piano* invokes thoughts about the sound of a piano. When we hear a sentence like (10), it evokes thoughts about how heavy the piano is. You can show that words that are related to the weight of the piano are relatively primed in the subject's memory in the second case. Words related to the sound of the piano are relatively primed in the first case. Now, when I say "evokes thoughts," I'm not actually saying that they are consciously thinking, "Oh my God, that's going to be heavy." What I mean is that, if I flashed the word *heavy* at you for 50 milliseconds, you would recognize the word more quickly then if I flashed some other unrelated word. The aspects of the meanings that are activated are shaded by the context in which a word occurs.

Implied Constituents. Another problem involves implied constituents. Let's take the following example:

(11) The man cut his steak.

It's quite likely that the man used a knife to cut his steak, and I think most people would agree than a knife is a constituent of the representation that we form of

this event. Linguists have argued such a thing from the fact that we can refer to that knife in the next sentence as though it had already been mentioned. We can say: "The man cut his steak. The knife was covered with strychnine. He died immediately." So that unmentioned argument is actually part of the representation that we expect people to form when listening to such sentences. Well, it's not in the sentence, so how could it be part of the meaning of the whole when it wasn't even in the sentence to start with?

(12) The boy ate.

Sentence (12) is a very simple case. He must have eaten something. I'd assume it was something that wasn't mentioned. So we can say, "The food was tasteless." This is the same point.

Content and Override. Maybe I will not trick you all on this one, but in psychological experiments when you ask people this question,

(13) How many animals of each kind did Moses take on the ark?

They typically say, "Two." And you ask, "Are you sure?" They say, "Yeah." You ask them to repeat the question. They say, "How many animals of each kind did Moses take on the ark?" You say, "Did Moses go on the ark?" And they say, "Moses? Moses? No! Moses didn't go on the ark." It so happens that the context here is so constraining both before and after the word *Moses* that it naturally leads us to understand *Noah* in this case. The context overrides the contribution to the meaning of the question that Moses would himself be expected to make. An interesting further fact is that, had the word *Nixon* been used here instead of *Moses*, it wouldn't have worked. Research has demonstrated this point (Erickson & Mattson, 1981). This tells us that something about this word actually influences the outcome. If the word was *Nixon* we wouldn't have understood *Noah*. But since *Moses* is a lot like *Noah* anyway, it's close enough, and almost all the clues lead to the same interpretation. Context can sometimes override the meaning of the individual words. By the way, the subjects in these studies are not people who do not know the difference between Moses and Noah. They are regular undergraduates who are easily tricked by sentences of this kind.

I do not think this finding is just a laboratory curiosity or a cute joke. It is actually an important element of everyday communication. If I say to my wife, who is on the other side of the room, near where my sneakers happen to be sitting, "Hand me my shoes, will you?" she won't say, "Those aren't shoes, they are sneakers." She'll just see what's there and hand them to me. If I say this to my 5-year-old, she may well say, "Daddy, those aren't your shoes,

they're your sneakers." So maybe its true that a 5-year-old operates on the compositionality of meaning principle. But as adults, we grow out of it!

Metonymy. Consider sentence (14):

(14) The ham sandwich needs his check.

I was made aware of this phenomenon by George Lakoff. A waitress can say this to another waitress. This means that the guy sitting at the table over there who ordered a ham sandwich is ready to leave and he needs his check. Now, obviously you can't get that so easily out of compositionality, because the ham sandwich isn't really what you're talking about. There is a very general device that we use in language. It's a way of adding a little color, I suppose you can say, or taking a shortcut when we talk; but it also has to be dealt with by any theory of understanding.

Metaphor. They symbol-processing theorists talk about metaphors as though they were atoms. For example, one can't understand *kick the bucket* in terms of the meanings of *kicked, the,* and *bucket,* because it's an atom. They support their claim by noting that, if you changed it around a little bit, it stops being that metaphor. You can't say, "Kicked over the bucket" and still mean, "The guy dropped dead." But, what about:

(15) His goose was cooked.

You can extend this in various ways. You can say things like, "His goose sure got cooked" or "He is stewing in his own juices now, isn't he?" You can draw on the semantics of the actual utterance and play out the metaphor, and anything that you can do structurally, rearranging it, making it a subordinate clause, all those kinds of things can be done to the metaphor also. So, it's not that structural rules don't imply internally to the metaphor. Yet, in some sense, the meaning of the whole is not given directly in the meanings in the individual parts.

(16) The haystack was important because the cloth ripped.

Context Dependence of Sentence Meaning. Example (16) shows that, in some cases, the compositional hypothesis breaks down entirely. Does anybody understand this sentence? It is sort of incomprehensible. You can say yes, but what it means is that, well, this cloth ripped and somehow that made this haystack become important. But what does it really mean? Think of a parachutist! When you have a critical piece of context, you can suddenly envision the scenario that this sentence is providing extra information about, and you can interpret what it means. But its meaning is really strongly dependent upon my having given you that scenario in terms of this poor parachutist whose parachute ripped. Now you can see that the meaning simply cannot be derived strictly from the sentence itself. In many cases, the meaning does depend upon the context in which the sentence occurs.

The examples we have been considering present a long litany of problems with the notion that you could get the meanings of sentences out of a procedure that uses rules that refer only to the form of the sentence and compose the meanings of the parts together to make a whole. Let me outline what's really happening here. This is a parallel argument to the one I gave you in the beginning. Speakers construct sentences to convey ideas that they have. The words in the sentences provide clues or hints that listeners use in conjunction with contexts and prior knowledge. Structure also provides clues indicating how to interpret the clues provided by the word. Sometimes the meaning of a sentence is just what we'd expect from compositionality, as in "John loves the girl" and "The girl loves John." Arguably, these are cases in point. But compositionality generally does not hold. What's generally the case is that the words give clues to the meaning of the sentence as a whole, and that those clues constrain, but are not themselves constituents of, the meaning of the sentence.

CONNECTIONIST MODELING

Now let's consider a view that doesn't start with symbols and atoms. Connectionist models have the characteristic of being "constraint satisfaction systems." Each piece of input, let's say each word in a sentence, can be thought of as exerting constraints on the interpretation of the sentences as a whole. A word sets up a pattern of activation over a collection of simple processing units that then exert influences on the patterns of activations in other units. As all these sets of units start to influence each other, the whole network will tend to settle into a stable state, and it's that state that represents the interpretation of the whole sentence. The individual words in the sentence do not so much have their meanings included in representations as parts, as they have their meanings acting as influences on where that constraint satisfaction process is going to end up. That's why a Connectionist approach to understanding language is likely to turn out to provide a better kind of answer to the problem of how people can figure out the meanings of the preceding examples than the principle of compositionality.

I will briefly describe a model that Mark St. John and I have developed (St. John & McClelland, in press). That model addresses five problems. The first one is the context sensitivity of the meanings of the words. For example, in "John threw the ball for charity," we have *ball* and *threw,* which are ambiguous words; they have their meanings determined by context. In "the container held the apples," we have the problem of specifying the meaning of a rather vague, general

term based on the particulars of the context in which it occurs. One of the reasons why the compositionalist hypothesis has tended to be appealing to people is that the Connectionists, especially I, have tended not to pay much attention up to now to structure, and to the fact that the ordering of the words and their present positioning and embedded clauses, and so on, influences interpretation. We have tended to think about this constraint satisfaction process as though each word was just another constraint. If there is a positive message in the symbol-processing approach, and some of the other criticisms of Connectionism, it is that we ought to show how we could really be more structure sensitive with our models.

Part of the motivation for this model is to show that we can get a network that follows this constraint satisfaction idea and is also structure sensitive. We want to use the structure of the sentence together with the words in it, to come up with the right meaning. We'd like to be able, for example, to get the model to understand who the agent is in a sentence like

(17) The bus driver gave the teacher the book.

and who is the recipient. We also want to deal with passive versions of similar sentences and to be able to use these kinds of structural characteristics and not just make guesses about semantics. So that's another goal of this model. The goal is to use the meaning together with the structure, not just the structure alone.

(18) The teacher ate the soup with the spoon.
(19) The teacher ate the soup with the bus driver.

We want to be able to know that, in the first sentence, the spoon is the instrument, and in the next sentence, the bus driver is not the instrument of eating, but is a co-agent (somebody who is participating in the act along with the agent). Therefore, the meaning is going to work together with the structure in this model. A fourth point, which is not as crucial in addressing the compositionalist hypotheses but which is important to me, is that language processing is a task that takes place in real time. We're always updating our interpretation of the sentence as a whole and with that we have anticipations as to what will come next. When somebody says,

(20) The landlord painted the walls with . . .

people expect something to come there, most likely *paint* or *Latex*. If the word is *cracks*, it actually takes people about a fifth of a second to incorporate that and adjust to the new interpretation, even though it is a perfectly plausible thing to say. We want our model to anticipate what hasn't yet been said based on what has already been said and to be able to absorb it quickly and easily if it fits; and

to be able to incorporate it, but with a little bit more difficulty, if it was not what was expected.

Here is our approach: we focus on a Connectionist learning principle to train a network to do the following task. We're going to give the network a representation of a sentence, and then we're going to give it a probe—that is, a question we could ask the network—and we are going to ask it to produce an answer. For example, we might have a representation of the sentence "The boy kicked the ball." The probe might ask for the agent. In this case, the answer is supposed to be "the boy."

If the network's answer is right, that's good. The network will be given feedback that tells it what the correct answer should be, and to the extent that it gets the right answer, things are fine. To the extent that the output mismatches the target, we will use that information to change the strengths of connections inside our network. We use a back propagation learning rule, which is basically driven by the mismatch between the output and the target.

The network has the basic architecture shown in Figure 7.2. There is a pool of units for representing the sentence. There is a pool of units for representing the probe. There is a pool of what we call *hidden units* and a pool of *output units* over which the answer that the network will give to our questions is displayed. The feedback will specify what those output units should have on them. In Figure 7.3 we see a close-up of part of the network shown in Figure 7.2. Let us consider, in particular, the probe units and the output units. Together these two sets of units allow us to query the network—that is, they allow us to put in questions and see what the net produces as an answer. In this case we've chosen to ask the network to respond to probes which represent questions like "What is the instrument of the sentence?" and "What role does the boy play?". This can be thought of as a matter of completing a role-filler pair where the probe is the role or the filler and the output is the completed pair.

To implement this, the probe and output pools each contain one unit for each role, and one unit for each concept. In the output we also have units for superordinate concepts. So, in response to *agent*, when the sentence is "The boy kissed the girl," the output should involve activation of the unit for agent and the units for boy, child, male, human, and so on, but not girl, adult, and so on.

Let us turn back now to Figure 7.2. The critical thing to note is that it is the representation on the sentence gestalt units there that allows us to complete each probe correctly. What are these representations like? We do not actually stipulate what these representations should be. Rather, we let them evolve through training. Once training has occurred (as described below), we imagine that these representations are formed, or activated, by the successive effects of the sequence of words in each sentence. The process is as follows. At the beginning of a sentence, the pattern of activation over the sentence gestalt is cleared, so that all units have zero activation. We present the constituents of the sentence one at a time on the input units. As each constituent is presented, it sends activation

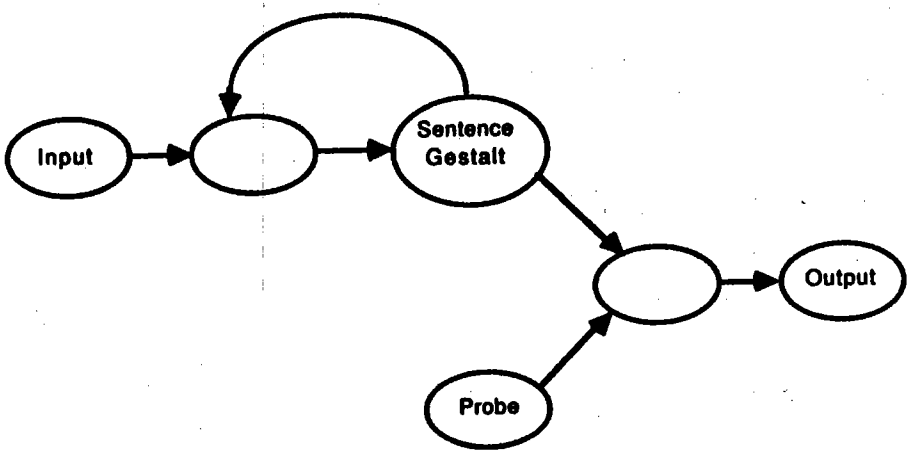

Figure 7.2. A sketch of the present conception of the sentence comprehension mechanism. The ovals represent groups of units, and the arrows represent modifiable connections.

through a set of connections to the first set of hidden units, and these in turn pass activation on up to the sentence units. After the presentation of each constituent, the resulting pattern over the sentence units becomes available as a context in which the next constituent is presented. That context, together with the next constituent, results in the next version of the representation of the sentence as a whole. The process continues until the end of the sentence is reached.

What's crucial in this process model is that we don't stipulate in advance exactly what the units will stand for in the sentence representation. Rather, we use a Connectionist learning procedure called *back propagation* of error (Rumelhart, Hinton, & Williams, 1986) to adjust connection weights all the way back through the network. The error arises from the discrepancy between the output the network actually produces in response to the probe and the pattern that would represent the correct response to the probe in this sentence. Roughly, if the network correctly activates boy and no other concepts in response to the probe *agent* after processing the sentence "The boy hit the ball," there is no error, and no adjustments are made. If the activation of boy is only partial or zero, and/or if incorrect units are activated, then there would be adjustments.

The forward going arrows in Figure 7.2 represent whole arrays of connections from units of one layer to units of another layer, and we adjust all those connections. That causes the network to learn to represent the inputs at the sentence level so that it can get the answers to the questions right. Thus, the network is using the Connectionist learning procedure to train itself to correctly represent the content of the sentence, where by "correctly represent it" we mean to have a pattern which can be used to cause the network to get the right answer to

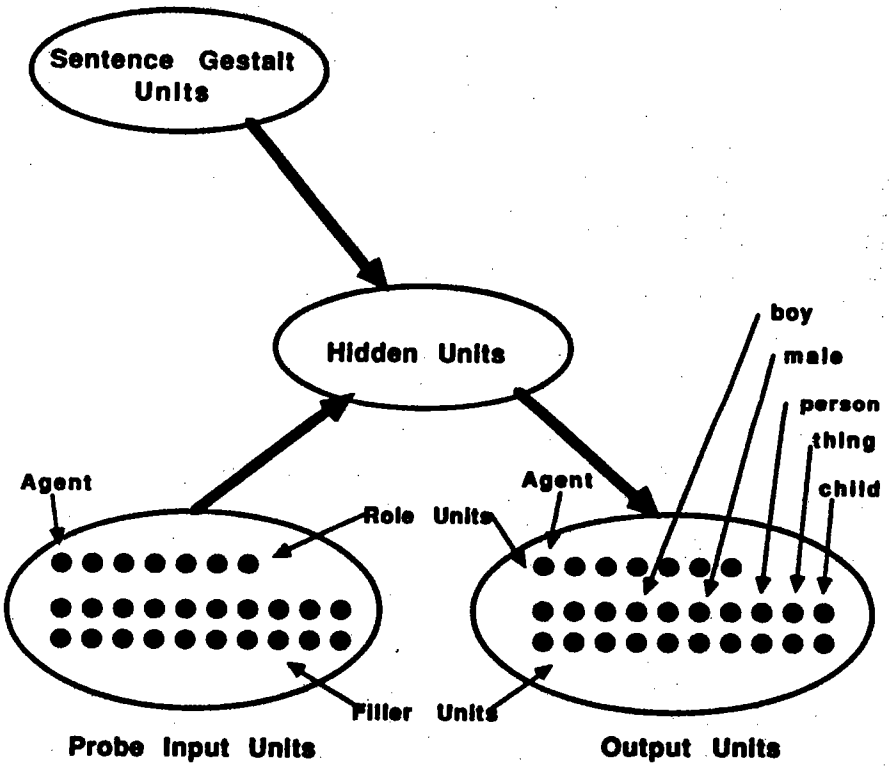

Figure 7.3. A close-up of the part of the network that is used for answering questions, illustrating some of the units in the "probe" and "output" pools.

questions. To do this, it must learn how to represent the different subjects and verbs and so on. What the network learns depends on what we train it on. While I will not go into all the details of the training process, there are a few things needed to understand the examples I am presenting. First of all, the network was trained with a fairly small corpus of words and possible events. Several of the words were words like *container* or *utensil* or *thing* or *adult*, which were vague in that they didn't specify particular detailed characteristics of the individual, they only limited it to a class. The rest of the content of the sentences tended to exert further constraints by virtue of the kinds of events that, say, men as opposed to women tended to participate in.

Here is one example. There was a bus driver who is a particular adult, and a teacher, another particular adult. The bus driver happened to be male. The teacher happened to be female. The bus driver tended to eat steak. The teacher tended to eat soup. As you see, this model has been subjected to certain stereotypes. These stereotypes reflect differences in the events that men and

women participate in. It is the same way as the other sorts of conceptual distinctions. The meanings of concepts are built up out of the contexts in which they occur. So human beings occur in one range of contexts, and the other kinds of animals occur in a different range of contexts. Similarly, males occur in one set of contexts and females in another: It is the same basic principle. Another important aspect of this model is that some of the constraints differ in strength. There is a tendency for the bus driver to be involved in eating steak as opposed to soup, but not absolutely. Bus drivers sometimes eat soup. On the other hand, other constraints are absolute. This bus driver also eats with gusto. He is never dainty. Whereas the teacher usually ate soup, sometimes she ate steak, but never with gusto. She was always dainty. So if you have the sentence,

(21) The adult ate the soup with daintiness.

it would tell you that the adult probably is the teacher. The idea, therefore, is that some constraints are strong, some are absolute. Constituents of events may be omitted from sentences describing those events. For example, I can say;

(22) The adult stirred the coffee.

and the network can still be asked whether he or she used a spoon or not, since the event that this sentence might be describing can actually involve somebody using a spoon. The network is responsible for answering questions about events even if the sentence doesn't contain all of the constituents of those events as stipulated in the words of the sentence. This is also a characteristic of the way we use language. We leave out the information that people can be expected to know based on what always happens in certain contexts. We don't always leave it out, but often we do. Some sentences use vague words, even though the event they describe involves a particular individual. In this model, there's always either the bus driver or the teacher eating either the soup or the steak. The model has to do its best to figure out what is the correct interpretation. The model is trained, and it learns rather slowly. One criticism of this model is that it takes a little bit too much computer time to get to the point where it knows how to answer the questions. However, once the model has learned, we can show that it actually does a reasonable job at meeting the goals that we set for it.

Meaning Disambiguation. Consider sentence (23):

(23) The pitcher hit the bat with the bat.

In the training corpus, this sentence was one where the constraints were such that the object of hitting was never a baseball bat. It's not likely that you hit the baseball bat with something. That never occurred in the corpus. However, a

flying bat was considered to be one of the things that could be hit by something. So, based on what the model saw in its experience, the first occurrence of *bat* in the sentence has got to be a flying bat. Similarly, the only kind of bat that you can use as an instrument, based on what the model saw, is a baseball bat. Furthermore, the word *pitcher* happens to be ambiguous. It could be either something that you pour things out of or a kid who's a pitcher. According to the model's scheme, the pitcher is a boy and he is a human being. After training, this sentence is interpreted by the model in the way that most of us would interpret it.

(24) The bus driver was given the rose by the teacher.

Structure Sensitivity. The above sentence is one of the hardest sentences that the model had to process. It's a passive dative sentence. After learning the corpus, we can probe with each concept to see what role the network thinks that concept fills. The network model's output is shown in Figure 7.4. In every case, the model is able to correctly assign each concept to its proper role. This sentence illustrates that the model is able to use the complex set of structural cues

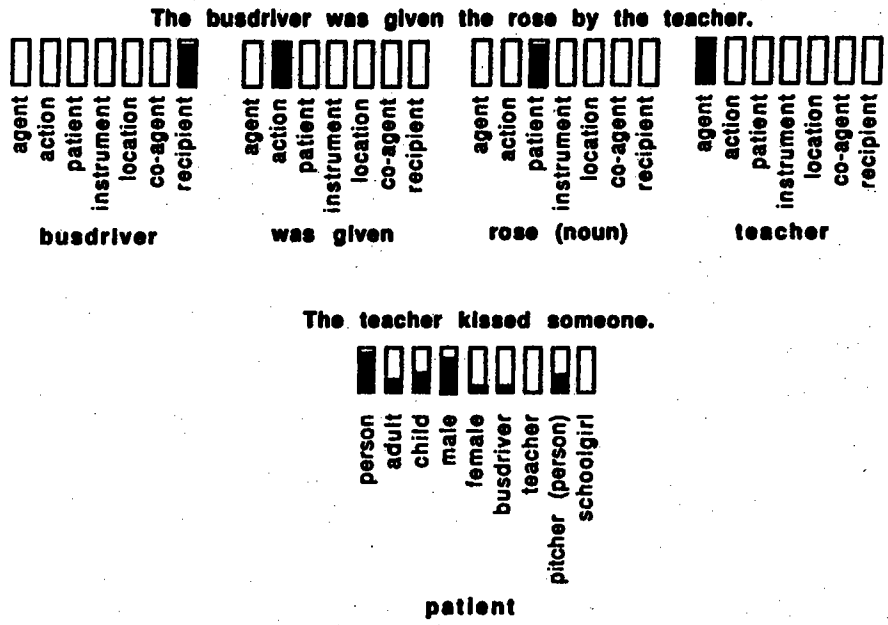

Figure 7.4. Activations of relevant output units in response to indicated probes after presentation of the sentence shown.

word order, verb, and preposition marking) to correctly assign words to the ppropriate roles.

This example shows that it is not the case that connections are always in- apable of being sensitive in that sentence.

Missing Arguments. If we give the model the sentence:

25) The teacher ate the soup.

ou can then ask it what the instrument is. It knows that the instrument is a poon. In the events that the model typically answers questions about, when omeone is eating soup the instrument is always a spoon. As a result, the model oesn't actually need to have the word *spoon* in the sentence in order to know iat a spoon is part of the event. This illustrates what I mean when I say that ords are providing constraints on the interpretation of the sentence as a whole. s interpretation contains information about the answers to questions about ings that weren't necessarily mentioned in the sentence, in a very natural and rect way. This process is the natural way in which interpretation is bound to ork when you take words as cues that constrain the meaning rather than as nstituents of meaning.

Instantiation of General Concepts. In this next example, we have cases here we use a general word like *something* or *someone*:

5) The school girls spread something with a knife.
7) The teacher kissed someone.

e only thing that was spreadable in the model's small world was jelly, so it ured out that it must have been jelly. In other cases the word *something* comes t to be something else. For example, consider the sentence (27). This sentence interesting, because it's not the case that the someone kissed is always male female. That depends on the sex of the subject. The model lives in a totally erosexual world, so its got to be a male person that is kissed in this case. And it is what the model reports (Figure 7.4).

Online Processing. I want to illustrate an online processing aspect of the del with the sentence:

) The adult ate the steak with daintiness.

s is not as relevant to the compositionality issue as it is to technical issues in nputational linguistics. The interesting problem is to figure out what is the st likely interpretation of *adult*. In order to do that, we have to use what is

called *right context*—that is, context that comes to the right of *adult* in the sentence. Sentence processing typically occurs from left to right. We want to read word by word and assign a meaning and then go on. The problem is that, often, we can't proceed in such a manner. We need to use constraints that occur later in the input. One way that this is done is by beam search. You set up all the possibilities as separately articulated cases and consider each one by following multiple paths. You also choose what seems to be the best option at the time, and if you get stuck, you backtrack. This model does something different. It has a pattern of activation which simply gets updated as each new constituent comes in. So it's a strict one-forward pass. The alternative interpretations are implicit in the single pattern of activation, thus eliminating the need to keep a record of all the alternative parts that are being considered.

We can see what's happening in Figure 7.5 if we look at the model's answer to the question "Who's the agent?" after each constituent in sentence (28). This

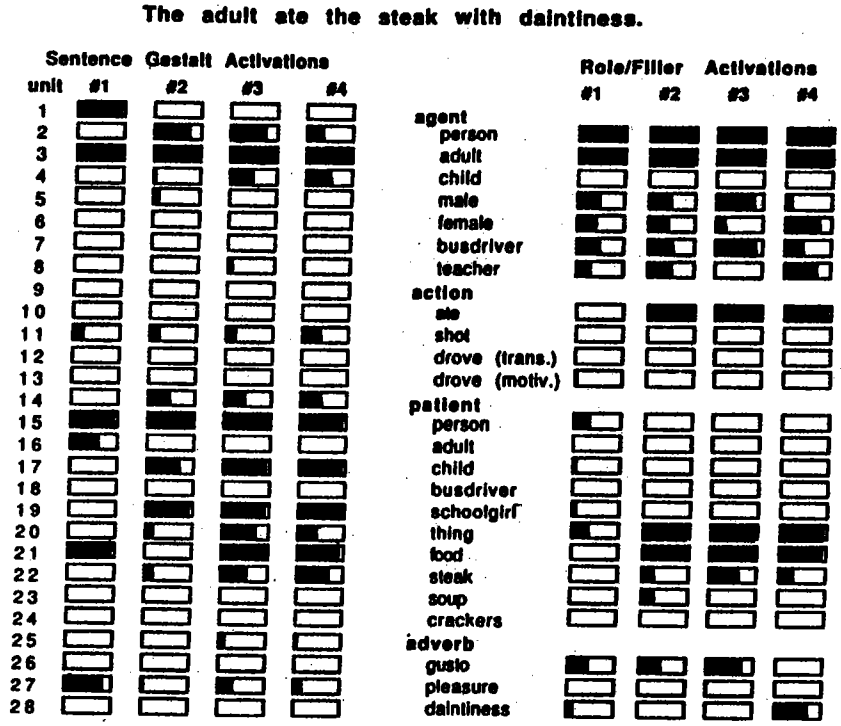

Figure 7.5. Activations of a subset of the sentence gestalt units (on the left) and of relevant output units in response to the indicated probes (on the right) after presentation of each constituent of the sentence "The adult ate the steak with daintiness."

information is present in the upper right portion of the figure, in the columns labeled #1 to #4. After *adult*, it has a slight bias towards male over female. Both units are partly activated. After *ate* it is totally neutral between male and female. *Bus driver* and *teacher* are both equally activated there. They are the two adults in this world, and thus far the sentence is neutral as to whether it is the male or the female. After *steak*, the model now feels that it must be the male, the bus driver. When you ask the question "Who's the agent?" it says, *male/bus driver* more than it says *female/teacher*. However, when we get the last constituent, "daintiness," which remember is never associated with the bus driver but is always associated with the teacher, the model reverses itself. The model was neutral at first, then favored one interpretation, and then reversed itself. This is all based on updating the representation as each new constituent comes in. Each word is exerting its effect as a constraint on the representation of the sentence as a whole.

I want to present one additional point about the inner workings of this model before I end. People like to think that words have meanings and I agree with that. If I say *wrist watch*, it sets one train of thought going, and if I say *fire alarm*, it starts a different one. Each word conveys something in and of itself across all the different context in which it occurs. In Figure 7.2, we can see what that invariant property is. It is the influence the word exerts on the first set of hidden units. Each word produces a set of influences on the activation of the hidden units which is the same every time that word occurs. The resulting activation depends on the prior context together with the word, that is, the activations of the hidden units are a joint effect of both the prior context and the word. But the influence which the word itself has is always the same. Now, we can look at the pattern of inputs each word generates and compare this for different words, to find out which words the model treats as similar to each other, and which ones it treats as different. To get a picture of this, we use a technique called *cluster analysis* that groups words according to similarity based on their input patterns each word produces.

The result of the analysis is shown for the verbs in Figure 7.6. For example, we can see that the patterns for the verbs *consumed* and *ate* are very similar to each other. The clustering analysis groups them by similarity and shows how similar they are. The more similar things are, the closer they are to the bottom of the graph. *Consumed* and *ate* are very similar; *drank* is quite similar to *consume* and *ate*, but slightly less similar than they are to each other. *Kissed, hit,* and *shot* are all similar to each other, perhaps because, as Drew McDermott of Yale once suggested, they're all words for "contact sports." *Stirred* and *spread* are similar, and *gave* which is the only dative verb that we actually used in this particular experiment, was represented quite differently from all the others. This analysis shows that the model assigns similar representations to words that share implications for the rest of the event as a whole. For example, *kissed, hit, & shot*

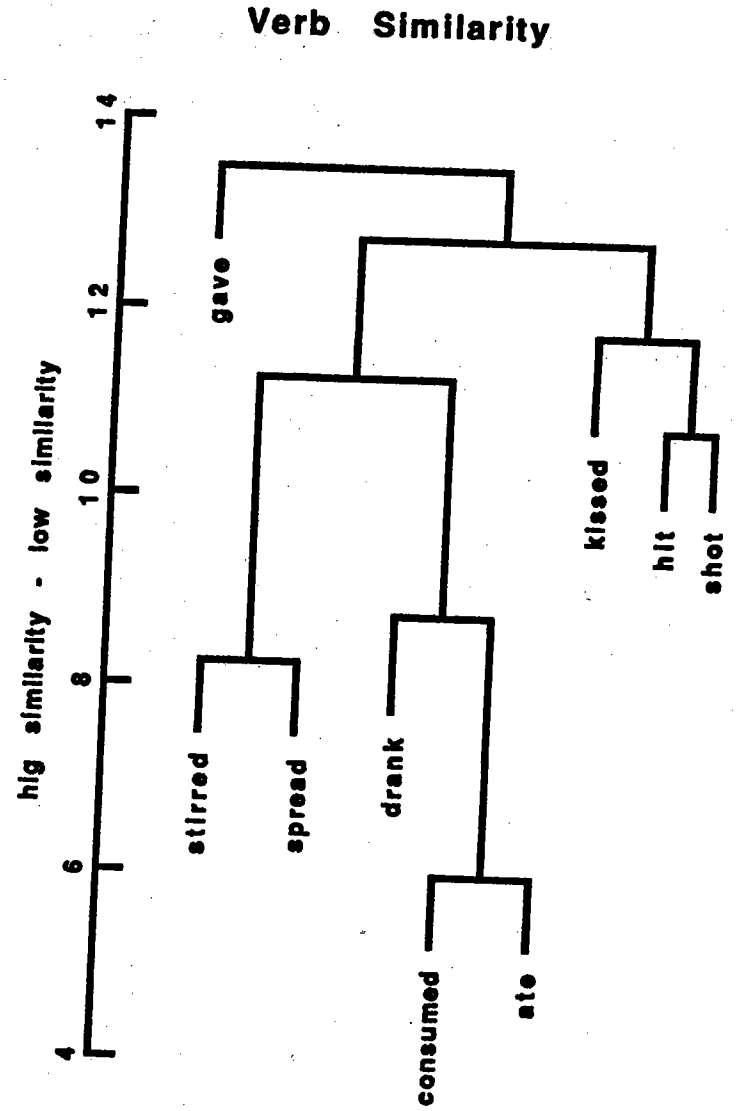

Figure 7.6. Cluster analysis of the weight vectors emanating from each word input unit to the hidden units in the comprehension part of the SG model, for the units representing the 11 unambiguous verbs shown. The vertical position of the horizontal bar joining two branches indicates the similarity of the leaves or branches joined.

all have animate subjects and they all have objects. *Hit* and *shot* can take either an animate or an inanimate object whereas *kissed* only takes an animate object. That would account for the difference between the relatively similar words *hit* and *shot*, and for *kissed* being a little bit more different. *Consumed*, *ate*, and *drank*, are similar, of course, because they have the same range of possible agents. *Ate* and *drank* both take objects, though these are different in the two cases; the events overlap in other cases as well. There is a small class of possible patients for those verbs. *Spread* and *stirred* have very similar argument structures to each other. And *gave* is, of course, quite different in that it has a recipient role as well as the object and the subject, and so there is a different set of constraints that it imposes compared to the other words.

We see, then, that, in the course of learning how to form these representations of sentences, the model has learned to let words that we might think of as being relatively similar have similar influences. So, in some sense, we've still captured something of the notion that there is something in common between the different occurrences of the same word. Yet we see that something not as a constituent of the meaning of the whole, but as a force operating in determining the direction in which the meaning of the whole is going to get pushed. But the eventual impact of this force depends on the context, as well as the word, as all of the phenomena I presented earlier indicate.

Let me summarize. Conventional symbolic approaches only partially characterize human language, because human language is not strictly compositional. An architecture is needed that can exploit the structure that conventional architectures can't. The stuff that compositionality doesn't seem to be well suited for is captured by parallel distributed processing models. These models are built out of neuronlike units and connections and trained by changing the connection weights. They provide mechanisms that can implement the constraint satisfaction process that seems to be required to go beyond compositionality. Many PDP models focus more on the structure that conventional architectures miss than on the stuff that they tend to capture. And the model I've presented is really no exception. I didn't deal with most of the more complex aspects of syntactic structure. In fact, I didn't deal with any kind of embedded structures. You may be asking yourself, "Well, OK, so they've begun to show some structure sensitivity as to word order within a single clause sentence. Are they going to be able to take the next step?" I'm afraid you will have to ask yourself that question for a couple of more weeks, because I don't know the answer yet. We've tended to focus our energy attempting to capture what the classical architecture has been leaving out as opposed to what they have been doing well. The reason is that I think the conventional architectures have reached their limit. There are certain things that they are good at, and that's been exploited. But there are problems in the area of language understanding things that they really have not successfully addressed. Our bet is that we need to start someplace a little different if we are going to go beyond these limitations.

SUGGESTIONS FOR FURTHER READING

Altmann, G. T. M. (Ed). (1989). *Parsing and interpretation.* Hove and London: Erlbaum.

Anderson, J. A., & Rosenfeld, E. (Eds.). (1988). *Neurocomputing: Foundations of research.* Cambridge, MA: The MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (2 Vols.). Cambridge, MA: Bradford Books.

REFERENCES

Chomsky, N. (1959). A review of verbal behavior by B. F. Skinner. *Language, 35,* 26-58.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior, 20,* 540-551.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 3-71.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 217-257.