CHAPTER 4

The Representation and Creation of Meanings

> "It is difficult to design and motivate empirical studies on concept acquisition without first committing oneself to a set of assumptions about what concepts are and how they are represented." (Keil, 1992, p.25)

4.1 Introduction

In chapter 2, we explored the nature of meanings and how concepts can be acquired, then in chapter 3 we investigated the particular problem of how learners can learn the meanings of unfamiliar words. In this chapter, these two strands will be linked to the wider field of evolutionary linguistics as discussed in chapter 1, as I take a look at recent simulations of the evolution of aspects of human language, and in particular at the models of meaning representation and meaning creation which have been put forward in the literature, and the model which I will adopt for the simulations in this thesis.

The linguistic competence of a language user falls naturally into three different, but mutually connected major subsystems: phonology describes the linguistic coding of the signals which are heard and uttered, semantics describes the coding of the meanings which are expressed and understood, and syntax can be regarded as the mapping between phonology and semantics. Although many linguistic theories choose to ignore or gloss over some of these subsystems, it is clear from the last four decades' work in linguistics that a comprehensive theory of language must address all three subsystems, as well as the interactions between the three. Keil's concerns in the epigram at the start of this chapter with respect to empirical studies with children are no less true when designing models for investigations using agents[1]. In order to be implemented on a computer, all three subsystems of language must be represented symbolically, and in different ways, so that they can then be interpreted by researchers as in some way ‘phonological’, ‘semantic’ or ‘syntactic’. The way in which these systems are implemented is heavily dependent on the theoretical assumptions of the designers of the simulations, and it is to these methods that I now turn my attention.

In this chapter, I explore in detail representations of meaning and mechanisms of meaning creation which have been put forward in evolutionary linguistic simulations, and then, building on the conclusions I draw from this, in chapter 5, I present my own model of semantic representation and meaning creation, which is used in the experiments in subsequent chapters. In more detail, section 4.2 is a discussion of the various semantic representations which have been used in recent simulations of aspects of language evolution in a little detail, discussing in particular how they relate to the famous dichotomy of meaning between *sense* and *reference* (Frege, 1892), and investigating the assumptions which have been made about how meanings are acquired and how they spread through a population of agents. In section 4.3, I then move on to look at the same simulations from the point of view of meaning creation, investigating the mechanisms which have been put forward to adapt internal semantic representations, and will suggest a suitable method for grounded, individually created semantic representations.

4.2 The Representation of Meanings

4.2.1 Predicate Logic

In the models of Kirby (2000, 2002), Hurford (2000) and Batali (2002), meanings are based on variant representations of first-order predicate logic, probably the most widely used knowledge representation language for describing the semantics of both simple propositions and fairly complex facts about the world which are derived from the simpler facts by standard formal rules of inference.

In the earliest of these evolutionary models, Kirby (2000)’s meanings each have three attributes, as shown in 4.1. He glosses them with standard linguistic theory as *agent*, *patient*, and *predicate*, while rightly emphasising that it is important to remember that

---

[1] In this thesis, I am not using the term *agent* in its usual linguistic sense of the logical subject of a transitive clause (see Song (2001) for an exposition of how the agent role is realised in different languages), but instead in its very common artificial intelligence sense, where it simply means a ‘simulated individual’. Under this umbrella term I include all simulated individuals, whether they exist only inside a computer or are physically implemented as robots.

these glosses do not exist in any way inside the simulations; they are simply a mnemonic which helps us to understand the structure of the meanings, which in effect are presented as a straightforward representation of a sentence containing a transitive verb with a subject and object. Kirby's three attributes are further classified, again in accordance with standard practice in this field, so that *agents* and *patients* are classed together as subsets of *objects*, while *predicates* are classed as *actions*. There is, therefore, a very precise typed structure to the meanings in Kirby (2000)'s semantic model; a particular object can appear in either of the two attributes which are available to it, namely as agent or patient[2], but it is impossible, for instance, for an action to occur as either agent or patient, or for an object to occur as the predicate.

$$\text{Meaning} = \; < \overbrace{\text{Agent} \quad , \quad \text{Patient}}^{\text{Objects}} \quad , \quad \overbrace{\text{Predicate}}^{\text{Actions}} > \tag{4.1}$$

In Kirby (2002)'s extension of this research, which focuses on the emergence of compositionality and recursion, the concepts are similar, as shown in 4.2, although there is a crucial extension. There are now two types of predicates: the first identical to that shown in 4.1; the second is a new type of predicate, which instead of an object as its second argument, takes another meaning representation. There are no further restrictions on the type of this embedded meaning: it can contain either a normal predicate or an embedding predicate, allowing in principle for unlimited recursion and an infinite number of meanings. This recursion is only possible, however, in the second argument position; only objects are allowed as the first argument to a predicate.

$$\text{Meaning} = \begin{cases} (\text{Predicate}_\alpha(\text{Object, Object})) \\ (\text{Predicate}_\beta(\text{Object, Meaning})) \end{cases} \tag{4.2}$$

Hurford (2000)'s semantic model is in a similar vein, based on a simple world of humans and animals, first described by Cann (1993). In Hurford's model, there is further expansion of the types of predicates which can be found, this time not just in terms of the type of patient they take, but also in terms of their valency, or the *number* of arguments they can take. In addition to the dyadic predicates which can be read as transitive verbs, as in Kirby's simulations described above, Hurford also has monadic relationships

---

[2] It seems that there is a further implicit restriction in the model, which ensures that the same object is never allowed to appear in both places of the predicate. For instance, there are no 'reflexive' meanings like <Agent=Zoltan, Patient=Zoltan, Predicate=Finds>

such as HAPPY and triadic relationships such as GIVE. Recursion is also implemented, but this time through not a whole class of embedding predicates, but just by one special SAY-predicate, which makes the further requirement that its agent must be human[3]. A description of Hurford's semantic model can be seen in 4.3.

$$(4.3) \qquad \text{Individual} = \begin{cases} \text{Human} \\ \text{Animal} \end{cases}$$

$$\text{Meaning} = \begin{cases} (\text{Predicate}_1(\text{Individual})) \\ (\text{Predicate}_2(\text{Individual, Individual})) \\ (\text{Predicate}_{\text{SAY}}(\text{Human, Meaning})) \\ (\text{Predicate}_3(\text{Individual, Individual, Individual})) \end{cases}$$

Batali (2002)'s semantic model differs slightly from those of Kirby and Hurford, in its use of variables, although the representations are clearly still based on predicate logic. Batali's representations are called *formula sets*, and are composed of a predicate and variables, or arguments, as shown in 4.4. Batali distinguishes two kinds of predicates, analogous to those shown in 4.3: monadic predicates, which he calls *properties*, and dyadic predicates, or *relations*. Two formula sets can be combined into another formula set by simply juxtaposing any number of them[4] (represented by the Kleene star notation in 4.4), and further manipulated by altering the mapping of the variables within them, to create more complex meanings. Batali deliberately chooses not to implement recursion directly, but nevertheless the repeated combination of formula sets produces in principle an infinite set of possible meanings.

$$(4.4) \qquad \text{Meaning} = \begin{cases} (\text{Predicate}_1 \quad x) \\ (\text{Predicate}_2 \quad x \quad y) \\ (\text{Meaning})^* \end{cases}$$

Having looked at the predicate logic representations used by Kirby, Hurford and Batali, it is interesting to investigate their semantic models in terms of their semantic content. In particular, what do the predicates and arguments *refer* to, and what *sense-relations* do the meanings have with each other? We might assume that the meaning of the predicates is

---

[3] All other predicates in Hurford appear to be able to take any individual, either human or animal, as any of their arguments.

[4] Batali has imposed an arbitrary limit of seven formula sets per meaning, for ease of implementation.

that of the common English words which are written identically to them, and that they are used to refer to actions in the model's imaginary world just as the English words are used to refer to actions or objects in the real world. But here we stumble across an important problem which recurs in many of these evolutionary simulations: the agents do *not* use the meanings to refer to actions or objects in their world, because there is no way in the experiment for the agents to access their world[5]. There is, therefore, nothing useful we can say about the reference of the predicates and arguments in these models; they have no denotation at all, because there is no external semantics in the models over which any denotation must be specified.

In order for us to be able to regard a meaning representation as encoding sense relations, at the very least there must be some structure in the representation, so that some relationship, however tenuous, between different elements (meanings) in the representation can exist. There are, therefore, some distinctions made in the models which we could arguably interpret as sense distinctions, particularly the hierarchical division of INDIVIDUAL into ANIMAL and HUMAN in Hurford (2000)'s simulations, which of course finds a parallel in the semantics of many natural languages. Crucially, however, we find that these 'sense' distinctions are *not* available to the agents in Hurford's model, who instead merely have two pre-defined, arbitrary classes of names, one of which can be used as an argument to any predicate, and the other which can be used as an argument to any predicate except SAY.

Overall, therefore, although each experimenter has implemented a structured representation which they have called 'semantic' in these models, there is very little about these representations which relates to sense and reference, and thus very little about them which can be sensibly regarded as in any way semantic, apart from the name itself. Instead, the purpose of the 'semantics' in these models is actually to serve as a blueprint for the syntax, which will then appear to emerge from the simulations. The agents' task is to learn a mapping between representations in two mediums: an existing, unchanging code which the experimenters call semantics, and a new, modifiable, emergent system which they call syntax. As Nehaniv (2000) has pointed out, syntax only develops successfully from unstructured signals in these cases because the signals are coupled with meanings which are already structured, and it is no coincidence that the emergent 'syntactic' structure directly parallels the pre-existing 'semantic' structure.

---

[5] Indeed, in Batali's model, there is no mention of an external world at all.

	a_0	a_1	a_2	a_3	a_4
b_0	a_0b_0	a_1b_0	a_2b_0	a_3b_0	a_4b_0
b_1	a_0b_1	a_1b_1	a_2b_1	a_3b_1	a_4b_1
b_2	a_0b_2	a_1b_2	a_2b_2	a_3b_2	a_4b_2
b_3	a_0b_3	a_1b_3	a_2b_3	a_3b_3	a_4b_3
b_4	a_0b_4	a_1b_4	a_2b_4	a_3b_4	a_4b_4

Table 4.1: Kirby (2001)'s model of meaning as a two-dimensional matrix with five discrete meanings on each axis.

4.2.2 Vectors and Matrices

The models discussed in this section use a different semantic representation,which is more abstract and less obviously based on a well-known formalism like predicate logic, but yet with meanings that still display a certain amount of the structure necessary for us to discern sense relations between meanings.

Kirby (2001) moves away from explicit predicate logic by introducing meanings which are vectors in two dimensions a and b[6]. Each dimension can range over a specified number of *discrete* values, and so the whole set of meanings, or the meaning space, can be thought of as a matrix, with a finite number of possible meanings, as can be seen in table 4.1, where there are 25 discrete meanings. Kirby (2001)'s model described above is very similar in its representation of meaning to one which was first presented by Steels (1996a). In this model, as in Kirby's, meanings are represented in terms of discrete values of features. Steels explicitly names both the features WEIGHT, SIZE, SHAPE and their respective values[7], but as in previous models, the names are merely mnemonics to help in understanding the model. The only real difference between the meaning space representations is merely that while Kirby's is a two-dimensional matrix with five discrete values on each dimension, Steels' is a three-dimensional matrix with three discrete values on each dimension. Kirby, therefore, has slightly reduced both the dimensionality and the number of possible meanings in the simulations, or cells in the matrix of meaning (25 (5^2) compared to 27 (3^3)), but otherwise the models' meaning representations are identical.

Brighton (2002), in a paper showing how compositional syntax arises under cultural pressures, extends the representations of both Steels (1996a) and Kirby (2001), by creating

---

[6] The two parts of the meaning could of course still be interpreted as predicate and argument, but this interpretation is no longer built in to the model.

[7] The possible values of the attributes WEIGHT, SIZE, SHAPE are { *oval, round, square* }, { *tall, small, medium* } and { *heavy, light, average* } respectively.

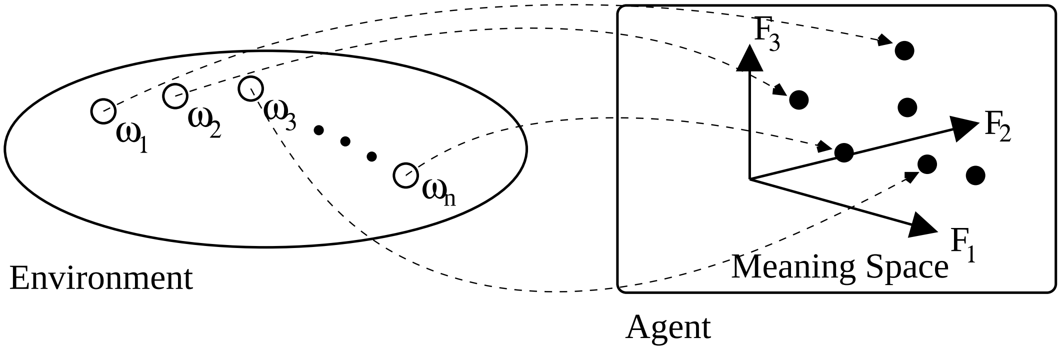

Figure 4.1: The relationship between the situations in the environment and points in the agent's meaning space, from Brighton (2002). The agent's meaning space is constructed as a multi-dimensional matrix, in F dimensions (here $F = 3$), with V discrete values possible on each dimension.

a more general meaning space, which is defined by two parameters: the number of features, or dimensions F, and the number of possible discrete values which can occur on each feature V, as shown in his diagram, which is reproduced as figure 4.1. Both the previous models, therefore, should be considered as specific instantiations of Brighton's more general model of meaning as a multi-dimensional matrix: Steels' can be defined with $F = 3$ and $V = 3$, while Kirby's can be defined with $F = 2$ and $V = 5$. It is important to point out, however, that figure 4.1 is potentially misleading in its depiction of the meaning space; despite its portrayal with apparently continuous axes, the meaning space is indeed constructed as a multi-dimensional matrix, with each dimension $F_1 \ldots F_F$ made up of a fixed, finite number (V) of *discrete* values.

Brighton (2002) also introduces an explicit external environment to the model, which consists of a number of *communicatively relevant situations*. These situations in the environment correspond in turn to distinct points in the discrete, multi-dimensional meaning space, as is portrayed by the dotted lines in figure 4.1. This mapping is specified randomly at the start of the simulation, and never changes thereafter. This representation of meaning as vectors clearly has a different underlying semantic model from the models of Kirby (2000) and Hurford (2000) discussed in the previous section. There is here an explicit external environment, and the meanings therefore appear to have *reference* to objects, or situations in this environment. The meaning space is also explicitly structured, so we can consider relationships between particular meanings, and it might be argued that the meanings do have some kind of *sense*, if we take a rather broad definition of sense as a relationship of any sort between meanings. For instance, meaning $\{1, 2, 2\}$ is related to meaning $\{2, 2, 2\}$ by virtue of the fact that it differs only in the first dimension, being identical in the second and third[8]. On the other hand, neither of these possible sources of semantic representation are quite what they seem; although the meanings appear to have reference, on closer inspection this turns out to be illusory, and although we do find some kind of relation which could be called a sense relation, this is not as great as might at first be envisaged by Brighton's notation.

Moreover, Brighton's generalisation algorithm itself is interesting, because of its great power based on extrapolating from chance correspondences to whole dimensions of meaning space on the basis of feature value identity and difference. We might imagine, for instance, that meaning $\{5, 1, 1\}$ could be considered as 'nearer' to meaning $\{3, 1, 1\}$ in terms of Euclidean distance than it is to meaning $\{8, 1, 1\}$, and therefore that distances relatively near to each other are more likely than distant ones to be considered as the 'same' meaning, but in fact this relationship is surprisingly not used in Brighton's model[9]. If an agent meets two meanings $\{5, 1, 1\}$ and $\{3, 1, 1\}$, both associated with the same signal, it does not use a simple generalisation, marking one signal with both meanings it has met (like $\{[35], 1, 1\}$[10]), nor does it even generalise across a contiguous portion of the meaning space, bounded by the meanings it has met (like $\{[3\text{-}5], 1, 1\}$), but actually it generalises across *all* possible meanings in the dimension where differences occurred ($\{?, 1, 1\}$), as shown in table 4.2. To take a real-world example of features with discrete values, let us imagine that the objects in Brighton's model represent chemical elements, and the first dimension represents the atomic number of the elements. When an agent meets two objects with the same signal, one of which is lithium (atomic number 3) and the other of which is boron (atomic number 5), Brighton's generaliser chooses not to mark the signal with a simple generalisation (*lithium* or *boron*), nor a spatial generalisation including the element which stands between lithium and boron in the periodic table (*lithium* or *beryllium* or *boron*), but generalises dimensionally across all elements, assuming that the atomic number, and the identity of the chemical element, is an irrelevant distinction for this signal. This is a legitimate, if very powerful, generalising strategy,

---

[8] This much, of course could also be said about the predicate logic representations discussed in the previous section: HAPPY(x) differs from HAPPY(y) only in its argument, as both expressions use the same predicate.

[9] In Brighton, Kirby, and Smith (2003)'s related model, on the other hand, the authors do indeed make use of this distance relationship in the meaning space to derive their measure of compositionality.

[10] The notation I use both in this paragraph and in table 4.2, is taken from the language of regular expressions (Friedl, 2002). In particular, I will make use of the following three expressions:

- $[xy]$ represents either x or y
- $[x\text{–}y]$ means either x or y or any other possible value between x and y
- and $?$ is a wildcard which matches any one possible value.

Table 4.2: Various methods of generalising over two meanings $\{5, 1, 1\}$ and $\{3, 1, 1\}$

Type	Notation	Members of Generalised Meaning
Simple	$\{[35], 1, 1\}$	$(\{3, 1, 1\}, \{5, 1, 1\})$
Spatial	$\{[3\text{-}5], 1, 1\}$	$(\{3, 1, 1\}, \{4, 1, 1\}, \{5, 1, 1\})$
Dimensional	$\{?, 1, 1\}$	$(\{1, 1, 1\} \ldots \{3, 1, 1\} \ldots \{5, 1, 1\} \ldots \{V, 1, 1\})$

which focuses on the similarities between two meanings and generalises over the differences, but it is important to note that Brighton's agents take account neither of any 'distance' between meanings, nor of how many possible meanings they are generalising over, and therefore that the number of meanings which this merged meaning $\{?, 1, 1\}$ corresponds to, and so the power of the whole generalising algorithm, are both explicitly determined by the particular value of V in each experiment. If V is relatively high, such as in the number of known chemical elements (currently 113)[11], then exposure to just two different values in one dimension causes the agent to assume that *all* different values of that dimension are expressed in the same way.

There are indeed some relationships between the meanings in Brighton (2002)'s model, which might charitably be interpreted as sense relations (although in truth they bear little resemblance to any traditional sense relations such as hyponymy and antonymy), but do these meanings have reference? The environment, and in particular its relationship to the agents' meaning representations, is not as important as it first seems in these models. Although the environment is explicitly linked to the meaning structure, by being defined as the source of the meanings, and represented as such in figure 4.1, on closer inspection we can see that the relationship between environment and meaning actually plays no role at all in the simulations; the agents never interact with the environment in any way, and the environment actually appears to be more of an obfuscatory factor in the model. We have seen in the previous section how the presence of an external environment is necessary for the development of a real semantic system, but now Brighton (2002)'s general model shows us that the *mere* presence of an environment is not enough: it is also necessary for the agents to have some interaction with their environment; without this, there is no way in which the meanings can have reference.

A direct extension of Steels' vector-based method of meaning representation is described by de Jong (2000), whose model is inspired by Cheney and Seyfarth (1990)'s study of vervet monkeys, and consequently whose agents' semantic 'state-action' space has three

---

[11] The apparent synthesis of element 118 has been retracted by Ninov et al. (2002), its purported discoverers.

state dimensions, representing the presence of a particular predator (S_1), and the agent's horizontal (S_2) and vertical (S_3) positions, together with two action dimensions, corresponding to movements horizontally (A_1) and vertically (A_2). The meanings described by de Jong, which he refers to as *situation concepts*, or patterns in the history of an agent's interaction with its environment, do not fit straightforwardly into Brighton's general model, because his meaning structure is not sufficiently uniform to be defined in simple terms with the two parameters F and V used by Brighton. Although the number of features (F) in de Jong is clearly five, the number of values on each features (V) is not uniform; after all, this model is tailored towards the specific problem of modelling the vervet communication system, rather than the more general problem of meaning creation.

For instance, the predator (S_1) feature has four possible values, representing the three specific predators and the absence of any of them. The other features fall naturally into two pairs, representing the horizontal (S_2 and A_1) and vertical (S_3 and A_2) positions, but each works slightly differently: The vertical positioning feature (S_3) has three possible values, and the vertical action feature (A_2), which defines a new vertical position for the agent, likewise has the same three possible values. The horizontal action feature (A_1), on the other hand, is represented explicitly in terms of movement relative to the current position: either to stay still, or to move one step to the left or to the right, again making three possible values. Because the horizontal action feature does not choose an absolute position, but instead defines its actions relative to the current position, the number of values on the horizontal position feature (S_2) is in principle unlimited[12], although in practice, when the predators appear in the world, they must be sufficiently near to the agents, in terms of their horizontal position, or else the agent's predator sensor does not detect them.

The model described by de Jong, nevertheless, has elements of both sense and reference relations in its meanings. The categories in the state-action space are related to each other using a hierarchical relationship, as we shall see in section 4.2.4, and they are also explicitly grounded in the agents' external world through the extraction of feature values.

4.2.3 Word Webs

Hashimoto (1997, 2001) presents a very different semantic model which is based on sense relationships between words. His focus is on the *sense-making* process and on

---

[12] In fact, the space used by de Jong (2000) is bounded, but I have been unable to find the limits to the horizontal plane which he used in the experiments.

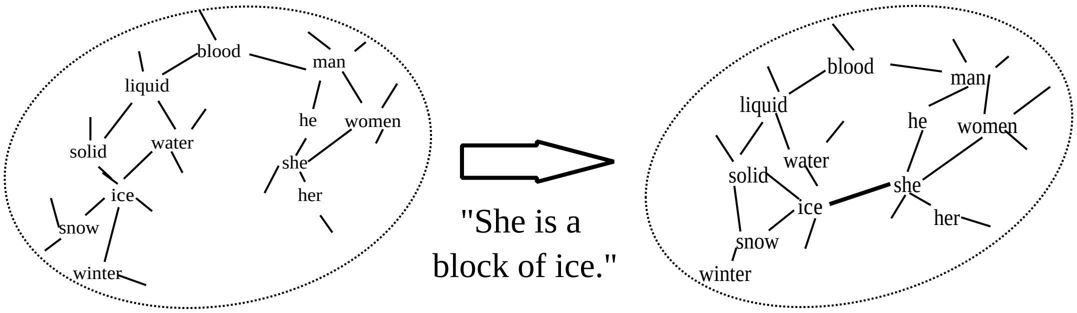

Figure 4.2: Semantic representation as a dynamic word-web, from Hashimoto (2001). The acceptance of the sentence by the agent triggers modification of its word-web.

language as a truly dynamic system, which is modified and remodelled after every communicative or linguistic episode. Interestingly, Hashimoto actually makes no distinction between words and word meanings, which are represented using an enormous word-web, implemented as a dynamic matrix, which models the relationships between words based on patterns of word usage and collocation in particular utterances and in larger texts of utterances, as shown in figure 4.2.

Hashimoto's semantic representation is clearly based on sense relations, although it is worth noting that the only relationship which is actually modelled is an amalgamation of word similarity (a measure of the frequency with which words are used in the same sentence) and word correlation (a measure of the patterns of word appearance in texts); there is again no modelling of even basic hierarchical sense relations, such as those we discussed in chapter 2. As a purely sense-based system, whose relationships are built from word usage patterns, we are not surprised to find that there is no reference at all in Hashimoto (2001)'s model. Again, there is no environment or world outside the agents, so there is no possibility that the words can refer to anything in this external world. This may also hold a clue to the lack of basic semantic notions such as *hyponymy* in this model; there is no way for the agents to discover that the set of referents referred to as CAT (its *extension*) is a subset of those referents referred to as ANIMAL, and in fact there is no way in which such a relationship can be represented in the basic word-web in figure 4.2, without further modifications which could potentially specify the type of the relationship represented by the connections between words.

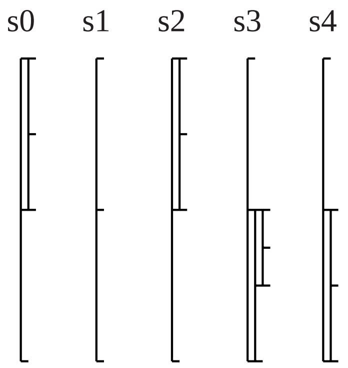

Figure 4.3: Steels' (1996) representation of meaning on discrimination trees. The discrimination trees are built on sensory channels (s0-s4), and are shown with the root of the tree at the left. Each segment of the tree shows the bounds between which it is sensitive.

4.2.4 Trees

In Steels (1996b), however, a different approach is put forward, which has been further developed in Steels (1997, 1999) and Steels and Kaplan (2002), and has been extended by many researchers since, including the work in this thesis: instead of defining a set of meanings which will be used by the agents in their language games, Steels simply defines a framework for representing meaning, on which the agents build their own individual representations. These semantic representations can be represented as a discrimination tree, with each segment showing the bounds between which it is active, as shown in figure 4.3. This semantic representation described by Steels can be clearly seen to have a reasonable number of sense relationships built into it; although not as comprehensive in its inclusion of multiple relationships between meanings as Hashimoto's word-web, Steels (1996b)' meanings have an obvious hierarchical structure, allowing the representation of semantic relationships such as hyponymy (one segment being a subset of another segment higher up the tree). As we saw in chapter 2, a binary tree structure also allows the implicit representation of antonymy, as each segment which has been refined into two subcategories necessarily has two co-hyponyms, which can each be regarded as the other one's antonym.

Steels' model is also closely bound to the environment in which the agents are situated, and, as we shall see in section 4.3, it is actually the main driving force behind the creation of meanings. Each segment on the tree, or category in the semantic representation, is abstract, and yet it also explicitly *refers* to a group of objects in the external world, namely those objects in the world whose feature values fall into the range to which the particular segment is sensitive. The bounds which define each category do not overlap, so the membership of a category is clear and distinct; for each feature value, at each level of the tree, there is only one branch which can be chosen, and only one possible meaning. Of course, because the tree is clearly hierarchical, the same feature value can still have different meanings at different levels of the tree, so a value which falls within the category defined by the upper quarter of a particular tree at the second level will automatically also fall into the category defined by the upper half of the tree at the first level[13], just as in our actual representation of meaning, the subcategory HERON is automatically also part of the larger category BIRD.

The meanings in de Jong (2000)'s model, which I discussed briefly in section 4.2.2, can also be thought of in terms of discrimination trees in a space, although the important difference between de Jong and Steels (1997) is that de Jong's meanings are each defined in all five dimensions of his meaning space at once. By contrast, although the agents in Steels' model do have multiple sensory channels on which discrimination trees are built, the channels are not related to each other multi-dimensionally (although segments on them can be combined to create compound meanings), and meanings are defined in one dimension alone, on each channel individually. Just as Steels' subcategories can be represented as ever smaller one-dimensional lines on the tree in figure 4.3, so de Jong's subcategories can be thought of as ever smaller five-dimensional subspaces on a multi-dimensional tree. Clearly, meanings which are represented as multi-dimensional subspaces are very difficult to represent graphically, and I will not try to do so here, but it is important to note that this multi-dimensionality of meaning has implications for the creation of meaning in de Jong's model, as I will discuss further in section 4.3.2.

We can see, therefore, that the meanings in Steels (1996b)' original model and its subsequent manifestations (Steels, 1997, 1999; Steels & Kaplan, 2002) and modifications (de Jong, 2000) clearly have both sense relations and reference relations, and are therefore the most truly semantic of any of the representations we have seen so far.

4.2.5 Prototypes

Vogt (2000), has implemented a model of meaning which bears some relation to de Jong's model, but with two main differences: it has been physically situated in actual robots, and the categories are the first to be based on the *prototype* model of meaning rather than the classical model, recalling our discussion in chapter 2. Vogt's categories are regions in a four-dimensional meaning space, and a particular meaning is defined by its relative

---

[13] Every value also falls into the category defined by the root of the tree, but this category is usually ignored, because it is of no practical use in helping the agents make sense of their world, as we shall see in section 4.3.

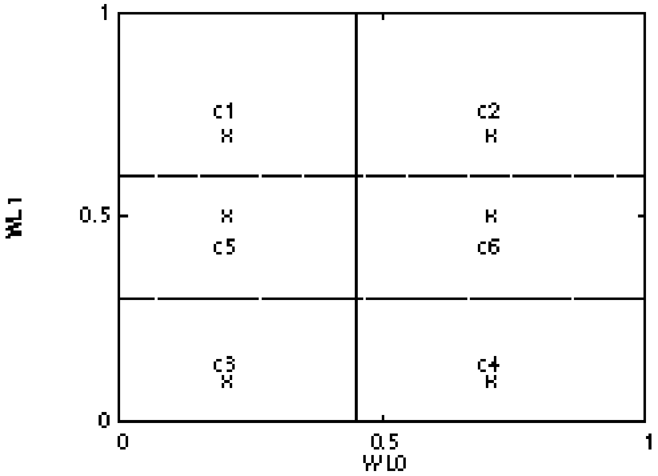

Figure 4.4: A representation of meaning as hyper-rectangles in a four-dimensional space, based on vicinity to prototypes, from Vogt (2000). Only two dimensions of the four-dimensional meaning space are shown, with the location of the prototypes marked by *(x)* and the names of the categories they form by *(c1-c6)*.

vicinity to one of the existing prototype points in the space, as shown in figure 4.4, which depicts just two of the four dimensions in order that the structure of the categories can be easily shown. The regions in Vogt's meaning space always have a hyper-rectangular shape, just as in de Jong (2000)'s model, so there is one important way in which Vogt's model of meaning necessarily deviates from an ideal prototype model; the boundaries between one category and another are clear and distinct, rather than fuzzy, making his model in this respect a compromise between a classical and a prototype representation. There is clearly some sense-like structure in the multi-dimensionality of both de Jong's and Vogt's meaning representations using subspaces, which is perhaps to be expected in structures which are derived explicitly from that of Steels (1996b). We can also see that the meanings represented inside both de Jong's computer agents and Vogt's physical robots have explicit reference to situations and objects which are encountered in the agents' environment.

Finally, I will investigate another different kind of meaning representation, in which meanings are again stored as prototypes. Despite the attractiveness of a prototype theory of meaning in certain situations, very few simulation models actually implement such a model, probably due to the difficulties involved in the representation of the system, and the concomitant processing power needed to run meaningful experiments. Belpaeme

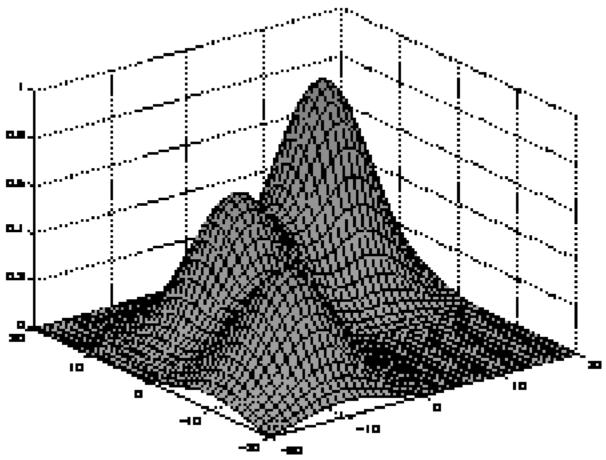

Figure 4.5: An adaptive network in a two-dimensional space, from Belpaeme (2002). This network, which represents one meaning, has three locally tuned units, each defined in terms of their centre, width and height.

(2002), however, in contrast to Vogt (2000), has managed to use a method of representing prototype meanings where the categories have fuzzy boundaries. He uses *adaptive networks* to represent categories, which are based on radial basis function networks (Orr, 1996), as shown in figure 4.5. Each category is represented by a different adaptive network, and each adaptive network is made up of a number of locally tuned units, which define the network. A locally tuned unit is defined by a Gaussian function around its centre; the function is always positive, but its value decreases monotonically as we move away from its centre, producing the characteristic bell-curves we see in figure 4.5. Each unit is also defined by its width, or the steepness of the curve's decline, and its weight, or the value of the function at the centre of the unit; the adaptive network in figure 4.5, which represents *one* meaning, has three locally tuned units with different centres and different weights, although each of the three functions has the same width and so the curves decline at the same rate.

The main advantage in Belpaeme's approach is that the meaning space is not divided into discrete regions, as in all the other approaches in which it makes sense to talk of a 'meaning space'. Instead, we can look at a point in meaning space in terms of the adaptive networks, by interpreting the value given by an adaptive network in response to a stimulus from the environment as a measure of category membership, or as a response to the question "how much of a [category name] is this stimulus?". The value produced by the adaptive network, therefore, naturally provides a fuzzy, graded notion of category membership consistent with that suggested by Rosch (1973). The main disadvantage is that the boundaries between categories barely exist at all, although they could of course be superimposed if required; an adaptive network gives some measure of category membership for every point in the space, so there is no way to say that a particular point is definitely *not-x*, and the boundary between categories cannot be clearly stated.

On the other hand, although Belpaeme (2002)'s model was clearly inspired by, and is primarily focused on, the evolution of colour categories in agents, his simulations actually have little to do with colours in particular and could have been presented as the evolution of any abstract categories; the stimuli received by his agents are essentially just points in a continuous three-dimensional space[14]. His model of meaning has a built-in measure of similarity between meanings based on the weighted sum of minimum distances for all the locally tuned units in each adaptive network, but does not lend itself easily to hierarchical or other sense relations. In common with the simulations we have looked at in the latter half of this chapter, however, the agent's categories are explicitly grounded in their environment, so they can be said to refer to stimuli or objects therein.

4.2.6 Summary

One of the crucial attributes which relates to the expression of reference is the idea that the agents have access to and are able to interact with some kind of external world, and objects therein which can be *referred* to. In the models whose meaning representation is based on predicate logic, such as Kirby (2000, 2002), Hurford (2000) and Batali (2002), this external world is missing, and the semantics presented is merely a code which the agents must decipher. In Steels' (1996a), Kirby (2001)'s and Brighton (2002)'s models, some structure has been added to the meaning representation, which is, as we have seen, a pre-requisite for the implementation of real semantic sense relations; additionally, both Steels and Brighton introduce into their models the notion of an external world, notwithstanding the fact that Brighton's external world is actually more of a distraction than an integral part of his model. Hashimoto (2001) presents a model which explicitly manages without reference, as it builds its semantic structure entirely on word collocations. Although no semantic notions other than collocation can be found, this model clearly has the potential to be extended to encode other semantic relationships reasonably straightforwardly.

By contrast, the meaning structures in Steels (1996b, 1997, 1999), Steels and Kaplan (2002) clearly contain both hierarchical sense relations and a real relationship with an outside world. It is not coincidental that these models are based not on providing an

---

[14] Belpaeme explicitly defines this space in terms of the L*a*b* space devised by the Commission Internationale de l'Eclairage, where L* represents lightness, a* red-greenness and b* yellow-blueness, but it is not clear that anything is gained by preferring this over a more abstract stimulus space.

innate set of meanings to the agents, but instead on enabling the agents to create their own meanings by providing them simply with a framework for representation and the ability to interact with their environment. Steels' original model has been extended to cover multi-dimensional meaning structures (de Jong, 2000; Vogt, 2000), and more substantially to incorporate representations of meanings as prototypes, with both discrete (Vogt, 2000) and fuzzy (Belpaeme, 2002) boundaries, without losing the properties of reference which are important to a semantic model.

It is clear that these latter, Steelsian, models are the most appropriate on which to build a model of meaning construction, and in chapters 6–9, I present a model based on this which will allow me to investigate how the interpretation of meaning affects the properties of agent-constructed communication systems.

4.3 The Creation of Meanings

In section 4.2, I surveyed many different systems for the representation of meaning in simulations which have been proposed by researchers into language evolution, and looked at how the conceptual systems relate to the Fregean notions of 'sense' and 'reference', which are often used to define meaning. Many types of meaning representation have been put forward, representing both sides of the divide between classical and prototype meanings we encountered in chapter 2, as well as more abstract representations based on predicate logic and mathematics. Having done this, we will now look at those same simulation models, but this time focusing on where the meanings originate and how they are created. Having already discussed the often acrimonious debate between nativists and empiricists, it is perhaps not too surprising to find a parallel, though altogether more amicable, dichotomy in the field of simulations of language evolution, between experimenters who provide a ready-made, 'innate' system of meaning for their agents on one hand, and those whose focus of enquiry is the creation of the meanings by the agents on the other.

In the first category fall the experimental models by Kirby (2000, 2001, 2002), Hurford (2000), Batali (2002), Brighton (2002), Brighton et al. (2003) and Hashimoto (1997, 2001), who all provide some kind of innate meaning representation for the agents at the start of the simulation. We can deal with these models briefly in this chapter, because the creation of meanings does not play a large role, if any, in their simulations. Typically, agents in the first group of models are provided with a finite set of meanings, according to whichever representation of meaning the experimenter has chosen, as I discussed in detail in section 4.2. During the experiments, the agents play two roles, with their exposure to these meanings being slightly different in each: as speakers, they are given a random meaning by the experimenter, which prompts them to produce an appropriate signal; as hearers, they receive both the speaker's signal and the meaning it expressed as a combined signal-meaning pair. The hearer's task is not to produce, but to try to discover the mapping between the two halves of the signal-meaning pair. There is no creation of meanings at all, therefore, unless we include the initial setup of the simulation, when a set of meanings is generated. If at any point new meanings are added to the agents' repertoire, then these too are explicitly generated and given to the agents. We can safely ignore such models, therefore, for the purposes of investigating the creation of meaning by the agents themselves.

On the other side of this particular ideological fence are the models by Steels (1996b, 1997, 1999), Steels and Kaplan (2002), de Jong (2000), Vogt (2000) and Belpaeme (2002); these experimenters provide the agents merely with the *capability* of creating meanings, and investigate the conditions under which they are successful. In these models, the development of the semantic space is an important part of the simulation, and so is much more interesting for our purposes. I will consider each of these models in turn, starting with those created by Steels (1996b, 1997, 1999), who created the basic framework from which all the others have been developed, and I will explore how the agents go about the process of developing their own semantic systems.

4.3.1 Discrimination Games

The basic procedure of agent-based grounded meaning creation, of agents developing meanings based on and relevant to the world they inhabit and the experiences they have, was initially modelled by Steels (1996b), who named it a *discrimination game*, after Wittgenstein (1953)'s famous *language games*. The Steelsian discrimination game is both selectionist, adaptive and minimalist: *selectionist* because the environment in which the game is played, and the dynamics of the game itself apply pressure to the agent's internal representations; *adaptive* because it responds to the results of the game to adapt its own internal representations in various ways; and *minimalist* because the agents in the simulations are provided only with basic operations for meaning creation, and not any intelligent generalisation or language-specific capabilities such as those which have been suggested for human infants and which we surveyed in chapter 3. I will briefly describe the four constituent parts of all discrimination games in the Steelsian paradigm, namely scene-setting, categorisation, discrimination, and adaptation, below, and will then go on to discuss its varying implementation by researchers, who each use slightly different methodologies, just as they used different methods of meaning representation.

scene-setting: the agent is given a specific discrimination task based on its environment, as follows:

- the agent is situated in a world made up of objects or situations, the features of which are in some way detectable by the agent.
- a set of objects or situations, called the context, is presented to the agent.
- one of the objects in the context is chosen to be the *target* of discrimination[15].

categorisation: the agent goes through all the objects in the context, returning for each an association with one or more of its existing semantic representations.

discrimination: the agent tries to find a distinctive category for the target. A category (or a set of categories) is distinctive if it is a valid representation of the target, and is not a valid representation of any other object in the context.

adaptation: the agent modifies its internal conceptual structure in some way; the methods of adaptation available to the agent are typically simple and few.

The processes of scene-setting and of discrimination itself are essentially fixed and identical in all implementations, although the Steelsian abstract model (Steels, 1996b, 1997) has been adapted into the *Talking Heads* experiments (Steels, 1999; Steels & Kaplan, 2002), in which real robots were built which could segment a scene into objects, and extract features from the scene they had developed, rather than being presented with the feature values from the objects. On the other hand, the particular methods of categorisation and adaptation of semantic representations are of course dependent on the particular semantic representation which has been adopted. In the next section, I shall briefly investigate the various implementations of the discrimination game.

4.3.2 Binary Category Splitting

The essential ingredients of the categorisation sub-task of the discrimination game are the reception of feature values from a space of possible values and the translation of these into a new space of possible categories. In Steels (1996b)'s model, and in his subsequent modifications thereof, including the implementation on the Talking Heads robots (Steels, 1997, 1999; Steels & Kaplan, 2002), the agents receive values from a number of different

---

[15] Steels (1996b) originally named this object the *topic* rather than target, but this term has connotations of conversational units, as well as a linguistic definition as "that element of a sentence which is presented as already existing in the discourse"(Trask, 1993), both of which can be misleading in a purely discriminatory situation.

features. In earlier models, the features were abstract without any specific meanings, but in the robot implementations, the features were pre-defined into spatial and colour characteristics such as AREA, HORIZONTAL and VERTICAL POSITION, HEIGHT, WIDTH, GREYNESS, RGB (the amount of red-, green- and blue-ness in an object), and the number of EDGES and ANGLES in the shape. For our purposes, however, the particular characteristics to which the features correspond are irrelevant, so I will generally regard the features as abstract, unless exemplification with a particular pre-defined characteristic is especially enlightening.

Each feature is independent as far as the agent is concerned, and the values it receives are normalised so that they always lie in the range range [0.0 . . . 1.0]. The translation from feature space to category space therefore involves a translation from an infinite number of possible values into a smaller number of categories (although also theoretically infinite). Steels' meaning representation is established on discrimination trees; the agent has a specific sensory channel for each feature, and on each sensory channel can build a separate discrimination tree. Each discrimination tree, therefore, corresponds to a specific feature of the objects in the world, underlying the conceptual independence of the features from each other. Each node on the tree is a category, and corresponds to a particular contiguous segment of the feature value space; the root node of the tree corresponds to the whole of the feature value space, i.e. it is bounded by 0.0 and 1.0 respectively. Categorisation, therefore, is the translation of a continuous feature value into a particular node on a discrimination tree. Of course, a category must exist before it can be used to categorise an object in terms of its feature value, and the agents adapt their semantic representation by one of the following means:

1. a new discrimination tree is created on a sensory channel which has no meaning structure.
2. an existing node on a discrimination tree is chosen, and meanings are added:
    - The region to which the existing node corresponds is split into two discrete segments, equal in size.
    - A new meaning is created for each of the new segments.
3. a node is pruned, or deleted, from the discrimination tree.

Because any created category can potentially be the source of a future refinement, the meanings created through this procedure fall naturally into the hierarchy shown in figure 4.6, which shows a very simple example of a discrimination tree which has been built

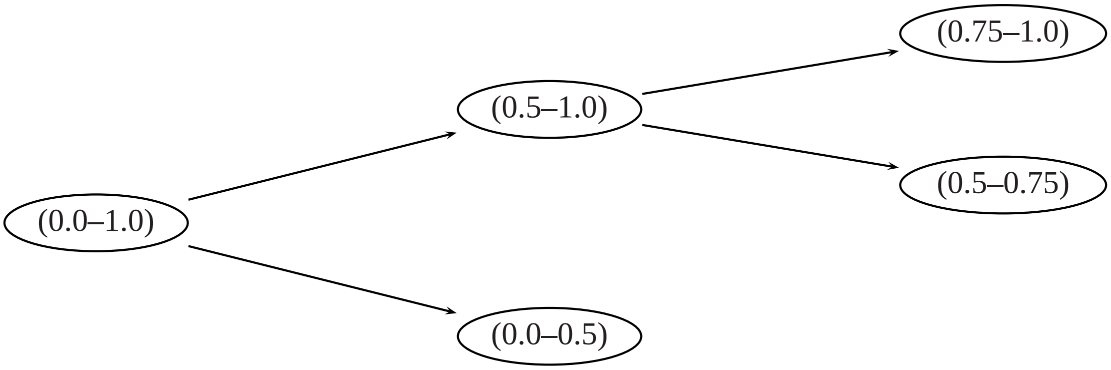

Figure 4.6: The development of categories on an abstract sensory channel shown as a discrimination tree. Each node on the tree shows the bounds between which it is sensitive; the root of the tree is sensitive to the entire feature value range (0.0–1.0); it has two daughter nodes, each of which is sensitive to half of the root node's range. The daughter nodes can also potentially have their own daughter nodes, and so the meanings can easily be represented in tree form.

on an abstract sensory channel. The root of the tree is sensitive to the entire feature value range (0.0 – 1.0); it has two daughter nodes, each of which is sensitive to half of the root node's range. In principle, a category could be divided into more than two segments, and the daughter categories need not have equal-sized sensitivity ranges, but as there is no limit to how fine-grained the distinctions which can be made even with the basic binary category splitting procedure, it seems sensible to stick initially with this framework, which is simple yet powerful, and ideal for exploring the development of meaning in agents. In Steels' models, failure in the discrimination game is the trigger for the adaptation of a sensory channel and the creation of more conceptual structure in the form of more specific categories. There is no pre-definition of which meanings should be created, however; the new categories *may* turn out to be useful in future discrimination games, but there is no guarantee. The agents in this model, therefore, have a mechanism for constructing concepts which are grounded in the environment (Harnad, 1990) and adaptive to their surroundings.

In de Jong (2000)'s models, the agents again receive information on sensory channels, but the feature values are specifically tied to particular representations of the state of the world, the actions of the agent, and an evaluation of the appropriateness of the action, as we discussed in section 4.2 which gives more detail of the structure of de Jong's meaning space. The environment provides a high reward for a specific action in the presence of each of the three different predators, corresponding to the appropriate evasive action taken by the vervets (see table 2.1), and provides high rewards for all actions when there is no predator around. The agents in de Jong's model have a fixed task, which is to create a sufficiently detailed semantic structure to successfully identify the presence of the predators and take the appropriate action.

The categorisation and adaptation processes in de Jong (2000)'s model are very similar to those in Steels (1997), although they may not appear so on first contact. The agents use the same kind of categorisation, in that they check whether or not a particular value falls into a subspace of the overall space, and the categories form a set of discrete categories. Interestingly, however, the original feature values in de Jong's model are not continuous, but already discrete, so it is unclear why he needs to introduce a categoriser which converts continuous variables into discrete ones, except that this kind of categoriser is of course more general and can be more easily used with other problems. The adaptive subspace method, as de Jong calls his process of concept formation, works in the same way as Steels' refinement process, with two main differences.

Firstly, the meanings are created across the whole of the multi-dimensional meaning space; although the task is set up so that the five sensor dimensions provide specific information, the agents do not know about this specificity, and search for general k-dimensional subspaces in whatever space they are provided with. The actual process of concept formation, however, is the same: a hyper-rectangular category is split into two smaller hyper-rectangles of equal size.

Secondly, however, de Jong's agents do not split categories blindly, without regard for whether the new categories will be useful in future, but instead decide whether to split a category or not on the basis of the pre-defined evaluative rewards they receive for particular combinations of states and actions from the simulation itself, and as such they are guided explicitly by a *reinforcement learning* process, which is useful for this kind of fixed problem. An agent is always looking for ways of splitting its meaning structure into subspaces, or *situation concepts*, considering a potential split in each dimension of its state-action space in turn. This potential split, as in Steels' (1997) model, would bisect the particular dimension, resulting into two smaller subspaces, both hyper-rectangles half the size of the original space. The criterion for whether to actually make the split is based on whether there is a significant difference between the distributions of the rewards for the experiences in each of the two potential subspaces. Summarising briefly, once all five dimensions have been investigated, a split takes place, realising the potential subspaces, in the particular dimension for which there is the greatest difference, as long as this difference is above a pre-defined threshold. In this way, the agents adapt their meaning space more quickly to the environment than Steels' agents, who are blindly creating

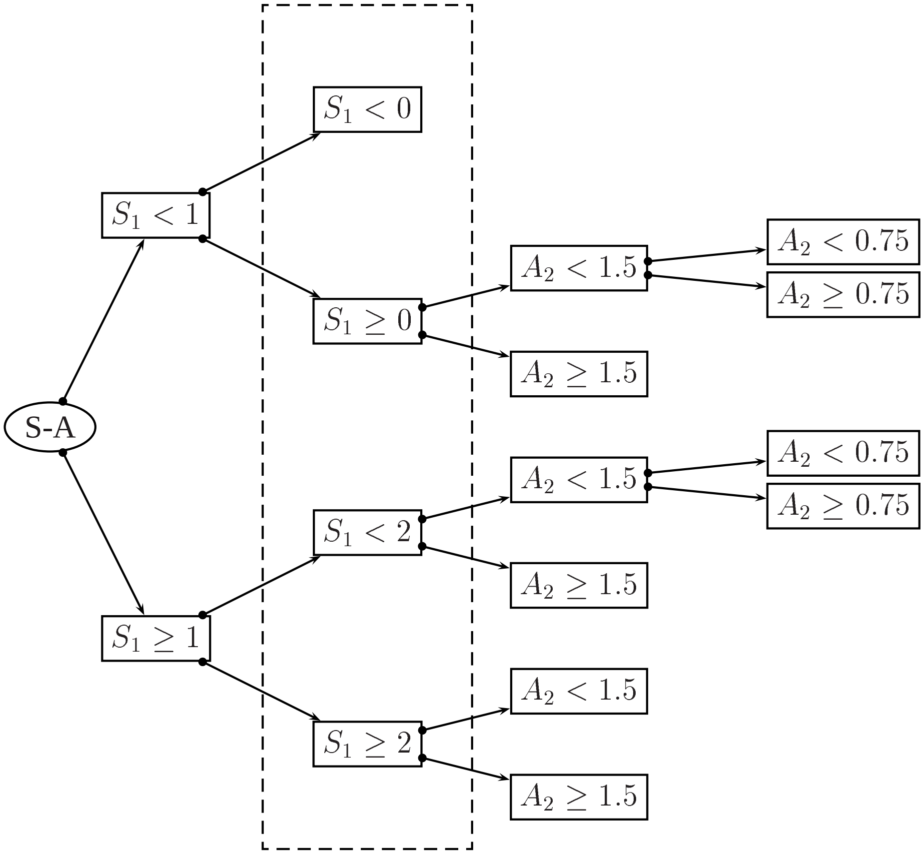

Figure 4.7: De Jong's representation of meaning in a five-dimensional hyperspace using adaptive subspaces. Distinctions have been made in the S_1 and A_2 dimensions.

categories which may or may not be useful to the agents in the future. In contrast, all the new categories created by de Jong's agents are necessarily useful in discriminating experiences; if they were not, then the split which created them would have remained a potential split, and would not have been confirmed.

Figure 4.7 is taken from de Jong (2000), and shows his representation of this meaning creation in five-dimensional space; again, only two of the dimensions are shown to make the figure comprehensible, because the particular task which the agents are set, to respond with appropriate actions in the presence of particular predators, can actually be solved without reference to the other dimensions. Recall that the dimensions of de Jong's meaning representation in a state-action space, as we saw in section 4.2.2, represent information both about the state of the world (S), and about the actions taken by the agents (A). In figure 4.7, the agent has split dimension S_1 into four situations, representing each of the three predators, and the 'safe' situation where no predator is found, which are all shown inside the dotted box in figure 4.7. For each of these situations, except the top one, which represents the safe situation, the agent has also split dimension A_2, which relates to the action it should take in terms of its own vertical dimension, and it has chosen a different action subspace for each of the three predator situations. In other words, figure 4.7 is a (fairly opaque) representation of the fact that the agent has successfully solved the task it was set, that is to find the appropriate action for each of the four different situations it finds itself in.

Although de Jong (2000)'s model solves the problem it was set, and expands Steels' (1997) original model into five dimensions, it remains problematic, on account of its design, which stems from the author's desire to associate the task so closely with the vervet monkey communication system we are now so familiar with. It seems to me that this kind of reinforcement learning paradigm is singularly unsuited to solving tasks of this nature; the agents in the model are not passing on innate knowledge like the vervets, but are being asked to learn from scratch the categories which are important to them, based on the feedback they get from the environment. This is all well and good, but if this was a real world consisting of vervet monkeys and predators, then the feedback they get from the environment would not allow them to solve the problem; rather the first time they chose the wrong action, they would be caught and killed. Reinforcement learning, which is based explicitly on learning from your mistakes, is not particularly useful if the cost of failure is so catastrophic as death, as Oliphant (1999) has noted.

4.3.3 Prototype Manipulation

In this section, I will explore the processes of categorisation and adaptation in the systems which used a prototype model of meaning (Vogt, 2000; Belpaeme, 2002). In both cases, recall that the overall structure of the discrimination game remains broadly the same as that designed by Steels (1996b) (see section 4.3.1), but that their particular implementation of categorisation and semantic adaptation are of course different.

In Vogt (2000)'s model, which is implemented on physical robots rather than inside a computer, categorisation works in the same way as in the models just described; the agents find out the space into which a particular feature vector falls, and return the category which defines this space. It is important to differentiate in Vogt's model between the *prototype*, which is a single point in the meaning space, and the *category*, which is a region in the feature space containing those points which are near the prototype. Although the prototype is the basis of the category, the category itself always has a hyper-rectangular shape, rather than a hyper-spherical one, so that no point in the space falls outside categorisation, and the boundaries between categories are clear and distinct.

Vogt's agents, however, also have a number of different models, or layers, of the same feature space, so that they can have overlapping categories on different dimensions.

Failure in the discrimination of objects is again the trigger for the adaptation of the agent's semantic representation, and this is done by increasing the number of prototypes in the world, up to a maximum resolution of the feature space, which is defined arbitrarily before the experiment. When adding categories, the agent chooses a random feature of the target object which it failed to discriminate, and uses the values of this to double the number of prototypes in the feature space; the new prototypes differ from the existing ones only in their positioning with respect to the chosen feature. For instance, assuming a three-dimensional space, and prototypes already existing at (0.1,0.2,0.3) and (0.9,0.2,0.3), the agent chooses the second feature or dimension, for which the target object had a value of 0.6. Two new prototypes are therefore created at (0,1.0.6,0.3) and (0.9,0.6,0.3), in the same position as the existing ones, except in the second dimension. This has the same effect of splitting the feature space in half, producing categories as hyper-rectangular boxes; the growth of categories can likewise be displayed on a (multi-dimensional) discrimination tree.

In addition, Vogt (2000)'s agents also update their prototypes when they succeed in using a category during a communicative episode, which I will discuss further in chapter 6. This addition to the model means that the categories adapt not only to failure, but also to success; the categories are adapted by shifting them slowly towards the feature vector which was successfully used. In order for communicative success to trigger the adaptation of categories, the agents must receive *feedback* from the model which allows it to evaluate communication. As we have already seen, the existence of feedback to language learners is hotly disputed (Bowerman, 1988) and is therefore absent from the model of agent-constructed communication which I will present in chapters 6–9. In effect, then, the agents in Vogt's model are using an instance-based learning technique (Mitchell, 1996), creating new prototypes when discrimination fails, and supplementing this with reinforcement of successful meanings when communication succeeds.

Belpaeme (2002)'s model of categorisation is different from all the others we have seen, because it is based on fuzzy prototypes; this means that for any object or stimulus, a measure of categorisation is returned. Category membership is no longer a binary yes/no decision, but is a matter of degree. When playing the discrimination game, therefore, Belpaeme's agents choose not the category which matches an object, but the category which *best* matches the object. Thereafter, the procedure is similar to that which we have seen before; if the target object's category is different to the category of all the other objects in the context, then the game succeeds.

Failure triggers the adaptation of the network in one of the following ways:

1. If the agent has no categories, then one is created which describes the topic, consisting of a network with one locally tuned unit, centered on the sensory representation of the topic.
2. An existing category is adapted to better represent the topic.
3. A new category is created.

A pre-defined threshold value inside the agents determines whether a new category is created or whether an existing one is modified. Adaptation of a network uses much the same procedure as the creation of a new network: a new locally tuned unit is added which is centered on the topic's representation. Belpaeme also tunes existing categories when they are successful, so that they are more like the topic, and has the locally tuned units decay over time, so that unused units eventually drop out of the category definition.

4.4 Summary

In this chapter, I have reviewed many recent simulations of the evolution of language, paying particular attention to their models of meaning representation and meaning creation. All of these models claim to have a 'semantic' meaning space, yet on closer inspection, the majority of the models had categories which were innate, pre-specified by the experimenters themselves, and had no reference to any external world at all. Many of them, in addition, had no sense relations of even the most basic type; there were no relationships at all between one category and another, which were instead atomic, individual items. The meanings in these models are not under the control of the agents at all; only the experimenters themselves can create new meanings or delete obsolete ones, and the meanings can only appear or disappear from an agent's repertoire by 'magic'. In summary, the semantic models of many language evolution simulations are simply not semantic at all, but are instead merely a rudimentary coding system, which the agents in the experiments use as a template with which to decode items expressed in another medium, namely the signals.

On the other hand, there are a sizeable number of experimenters who have made an effort to incorporate some kind of realistic semantic systems, by including an external world of objects which the categories refer to, and by providing various different methods for creating meaning based on the agents' experience in this world. Meaning creation in these models is based on the *discrimination game*, in which an agent's task is to return a category which describes one particular object, distinguishing it from another set of objects. Experimenters have used both classical categorisation and prototype categorisation successfully for their semantic representations; both have advantages and disadvantages, but for the remainder of this thesis I will use my model of semantic representation and creation, which is explored in more detail in the next chapter.