A Model of the Acquisition of Spatial-Relation Concepts and Words

Eliana Colunga-Leal and Michael Gasser
Computer Science Department, Cognitive Science Program
Indiana University, Bloomington, IN 47405

Abstract

This paper is about the acquisition of spatial relations, both conceptual and linguistic. Spatial relations, we propose, are defined by correlations across different perceptual modalities. The nature of the correlations that define a relation determines how hard learning a word for that relation will be. We present a connectionist model that learns about relations in two phases, a pre-linguistic phase in which concepts are organized according to the perceptual input, and a linguistic phase, in which pre-existing concepts are re-organized under the influence of words. This model accounts for the order in which spatial relation terms are learned by children.

1 INTRODUCTION

When we look at a scene we identify objects, but we do more than that; we also identify the spatial relations between those objects. When we describe a scene, we do so by talking about the objects and the spatial relations between them. This may seem very natural for us, but children have to learn to do this. Many spatial relations are learned fairly early — words like *under* and *out* appear by 25 months of age — while others seem to be harder. An extreme case are the words *left* and *right*, which cause trouble even for adults. What is it that makes one spatial relation harder to learn than another?

To answer this question, we will concentrate on four spatial terms which we think illustrate the main points in our account: *on*, *under*, *left* and *right*. We first present a short review of the data available on this issue as well as several models that have been proposed to account for it. Then we present our own account — the differences in difficulty reside in the pattern of correlations in the input, linguistic and non-linguistic — together with the network that implements it. Finally we present simulations that mirror some of the human data.

2 THE PROBLEM

Intuitively, it seems clear that the arrangements shown in the four parts of Figure 1 depict different spatial relations. One could argue that since they look different, they will produce different patterns of activation across the retina and ultimately different higher-level representations. Thinking of this, one would not expect ON and UNDER to be any easier to learn than LEFT and RIGHT. In a sense, this is true, as as shown by recent evidence suggesting that infants as young as 3–4 months are able to categorize both LEFT-RIGHT (Behl-Chadha, 1995) and ABOVE-BELOW (Quinn et al., 1996) spatial relations. But this is not all there is to spatial relations.

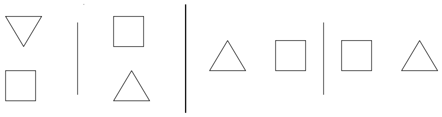

Figure 1: Some simple spatial relations

Categorizing spatial relations goes beyond visual input. There is more to *on* than (VERTICALLY_ORIENTED$(x, y) \wedge$ CONTACT(x, y)) There is also the experience of putting and watching things being put on other things and of being on things. All this, we think, is necessary to develop the notion of SUPPORT that seems to be at the heart of understanding *on*.

Evidence from language acquisition agrees with this; spatial-relation words are not all equally hard. There is a general pattern that babies follow when learning spatial relations. *Under* and *on top of* are among the first relational terms learned by children, being comprehended at around 16 months and produced by 30 months of age (Fenson et al., 1993). *Next to* and *beside* appear a little later and *to the left of* and *to the right of* much later (Clark, 1973).

This difficulty with *left* and *right* is further supported by studies showing that 4-year-old children have trouble performing left-right discriminations (Rudel, 1963). For a long time, it was thought that left-right discrimination problems were due to lack of maturation, but recent evidence shows that 4-year-olds can be trained to solve the task (Braine, 1988).

Gibson (1969) accounts for them by positing attentional defficits. Braine & Fisher (1988) claim they occur because of insufficient cognitive development. Logan (1995) suggests they are a result of a failure to maintain the different frames of reference that are necessary to solve the task. Clark (1973) suggests this difficulty arises from the lack of supportive bodily asymmetries for *left* and *right*.

We agree with Clark. In our view, the difficuly in learning the different spatial relations depends only on the nature of the correlations that define a given relation. Our model accounts for both the difficulty in learning *left* and *right* and the ability of young infants to categorize LEFT–RIGHT spatial relations as readily as ON–UNDER relations.

3 THE MODEL

The model we will be describing is part of a larger project, Playpen, whose goal is to model the development of spatial cognition (Figure 1). We are particularly interested in how vision, motion, proprioception and language interact in shaping spatial concepts throughout development (Gasser and Colunga-Leal, 1997). Here we will concentrate mostly on the Where side of the visual system (Figure 1b) because that is where categorical spatial relations are formed (Kosslyn, 1994).

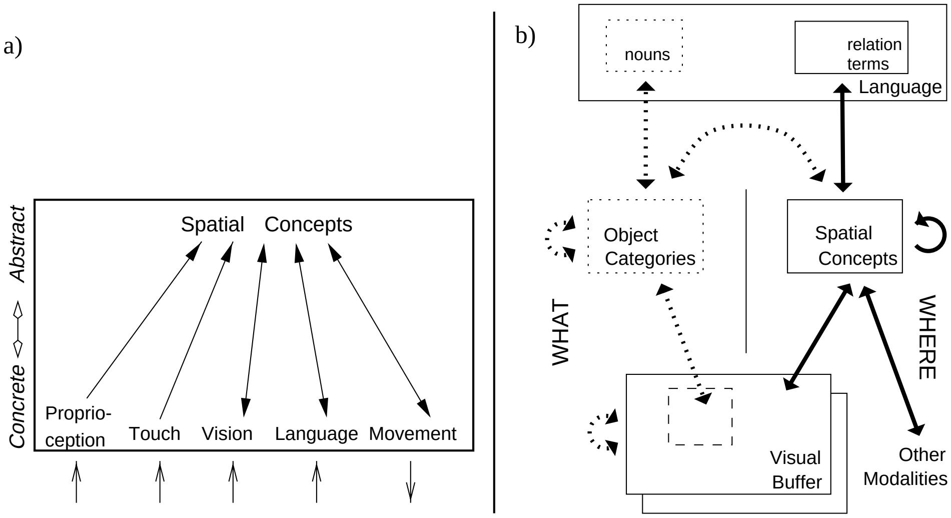

Figure 2: a) Playpen architecture b) a closer look at Playpen's visual system

Our model is organized as a series of maps at different levels of abstraction (Kosslyn, 1994). At the lowest level, input from the senses is stored very much like it is perceived. As we move up, representations get more abstract, starting from specific instances and going to categories that are more or less context-free. At the highest level, maps from different domains are associated with one another (Edelman, 1987), resulting in the highest level of abstraction.

According to our model, at the lowest level all spatial relations are equally hard. However, as we move up, some relations are more easily clumped into categories. This is because of the way the different domains that define the relation correlate. The more and better the correlations, the easier it is to abstract it. For example, if two events correlate in several modalities, they will be more similar, leading the higher level, which can relate different modalities, to cluster them together. This behavior allows the system to make better predictions of what happens in the world. As the number of correlations associated with a category increases, so does its predictability. For example, if things that make good supporters have a flat surface, knowing that a "dax" is on a "blicket" will lead you to believe that the blicket has a flat surface. By the same token, if you know something has a flat surface, you know things can be put on it.

We need to be able to represent relational knowledge of this sort. In language, spatial relations refer to (usually) two things that are being related to one another. The position of the noun phrases phrases referring to these things matters: there is a place for the **trajector**, the thing that is being related and a place for the **landmark**, the thing the trajector is being related to. The trajector appears to match the perceptually defined **figure**

The problem with *left* and *right* seems to be tied to language; it is not purely perceptual. All the evidence suggesting that left and right are hard involves labeling of some kind, while evidence for how they are easy involves no naming. This problem could come from the speaker's and the listener's failure to agree on what the trajector of a scene should be. If this happens, the listener is getting the *wrong* input.

Here, again, correlations help sort things out. Relations occur in a context and each context brings different things to mind. For example in an ON situation some aspects that are important for an UNDER situation, such as whether the thing on the bottom is completely covered by the thing on the top, may not be noticed. In an UNDER situation, contact, which is very important for ON, may be ignored. This regularities are exploited by the learning mechanism.

Learning occurs in two phases. We believe that even before language comes into the picture, a child is trying to make sense out of the world. It is to the child's advantage to be able to predict the outcome of his own actions as well as of the actions of other objects in the world. This is not done in a conscious way; the child is merely storing events as experienced through the different sensory modalities. However, there is structure in these events, there are correlations, and so during this pre-linguistic phase the first spatial concepts appear, representing highly correlated events in a more efficient way.

The second phase occurs when language comes in. Language works on the spatial concepts formed during the previous phase; hence, words that make a better match for already existing concepts are more readily learned that those that do not. At the same time, language modifies the existing concepts; since language is just another input, it provides new correlations which may be capitalized on to aid in the formation of useful categories.

In sum, this accounts for the relative difficulty of learning *on (top of)* and *under* and *left* and *right*. In the case of *on*, there are characteristics of the objects involved in the relation that occur frequently in ON situations; for example, the object on the top tends to be smaller and the object on the bottom tends to have a flat top. There are also asymmetries that are specific to the relation. For example, an ON situation is the sort of situation that occurs from stacking things up, the object on the top was the last to move and is the most likely to go away, removing the object at the bottom makes the one on top fall, and so on. Even more, these asymmetries correlate nicely with asymmetries caused by gravity in other domains such as proprioception: it takes more effort to lift things higher, being lifted feels different than being put down. So instances of ON are *similar* to each other. *Under*, on the other hand, is typically applied to other situations, those in which one object is covered, and often obscured, by another. Instance of UNDER are similar to each but different from instances of ON.

On the other hand, LEFT/RIGHT has little to work with. There is no clue in size as to whether the object is on the right or on the left, they are both equally likely to go away, and either of them could have gotten there first. There is also the problem that as one moves with respect to a horizontal relation, the objects involved switch places. All this resuls in more variability between instances of LEFT/RIGHT situations and fewer correlated asymmetries to distinguish between LEFT and RIGHT.

A further consequence of these correlation patterns is that having a category that encompasses both LEFT and RIGHT (such as *next to*) makes sense because there are few asymmetric dimensions to distinguish the relations and, at the same time, some shared dimensions that bring them together. For example, pushing one of the objects may move the other. This means that before language appears, ON and UNDER will have a reason to be separated, whereas LEFT and RIGHT will not. It is only when language comes in that LEFT and RIGHT will be pulled apart.

4 THE NETWORK

The portion of the Playpen architecture which is relevant for the simulations discussed here is shown in Figure 2b. The network is of the generalized Hopfield type: connections between units are symmetric, and units repeatedly update until the network settles. The network is made up of two kinds of units: **Object Units** (OU) and **Relation Units**(RU). OUs are oscillators: each has a relative phase angle in addition to an activation. As in a number of other recent models (Hummel and Biederman, 1992; Shastri and Ajjanagadde, 1993), phase angle functions to bind together the features of distinct objects. Units with the same phase angle are part of the same object, and units with different phase angles belong to different objects. The connection between each pair of OUs has not only a weight but also an associated **coupling function**, a function of the difference in phase angles of the two units. The coupling function must be symmetric about 0 and its derivative derivative must be anti-symmetric about 0. Both the activation and the phase angle of an OU are potentially modified each time a unit is updated. The input h and change in phase angle $\Delta\phi$ to an OU i are given by

$$h_i = \sum_{j=1}^{n} a_j \cdot w_{ij} \cdot \Phi_{ij}(\phi_i - \phi_j) \tag{1}$$

$$\Delta\phi_i = \frac{\pi}{\sum_{j=1}^{n}} \sum_{j=1}^{n} a_j \cdot w_{ij} \cdot \Phi'_{ij}(\phi_i - \phi_j), \tag{2}$$

where n is number of units in the network, a_j is the activation of unit j, w_{ij} is the weight connecting units i and j, and Φ_{ij} is the coupling function associated with units i and j. A stable state of the network is, then, a state in which both activations and phase angles are not changing.

RUs, which are the major innovation of the architecture, are used to represent relational information; they are "about" two different objects. Each RU is made up of a cluster of simple object units hard-wired in such a way that the unit as a whole is activated to the extent that it is receiving input from two distinct objects. It has two **interfaces**, each consisting of pairs of OUs: one to handle interaction with other RUs, the other to handle interaction with OUs. The hard-wired connections within an RU try to align the phase angles of the corresponding units in the two interfaces. An RU is considered activated when all four of its interface units are activated.

The input to the network is of two types: linguistic and non-linguistic. Non-linguistic input is represented using OUs. The **Visual Buffer** is organized as a group of feature maps, each of which represents the presence or absence of a visual feature, such as color or texture, in space. The other kind of non-linguistic input is in the **Other Modalities** layers, a series of dimension-like groups of OUs. We assume these dimensions come from other non-visual modalities such as proprioception and from combinations of visual and other information, but we currently do not model the process by which the dimensions are abstracted out of lower-level input.

Linguistic input is presented to the **Language** layer. RUs are used to represent relational terms. Within each Language RU, one unit within each of the interface pairs is designated the trajector, the other as the landmark of the relation. An activated Language RU represents a particular relation term as well as phase angle assignment to the objects which represent the trajector and landmark of the relation. Language RU are connected to each other and to RUs in the Spatial Concepts layer.

The **Spatial Concepts** layer also consists of RUs. RUs in this layer represent micro-relations, and spatial relations in the network take the form of patterns of activation across these units as well as the Language-level RUs. Each Spatial Concepts RU is connected to a pair of non-linguistic input OUs and to all of the RUs in the Language layer. During pre-linguistic learning, the pattern of connections which develops between the Visual Buffer and Spatial Concepts layers and within the Spatial Concepts layer represents the network's pre-linguistic understanding of spatial relations, reflecting the correlations found within and across the input domains. During linguistic learning, the addition of Language-level inputs and outputs modifies the system's spatial relations, as connection weights develop between the Spatial Concepts and Language layer and within the Language layer and as other connection weights are modified.

The network is trained using a variant of Contrastive Hebbian Learning (Movellan, 1990), modified to accommodate unsupervised learning (auto-association) and phase angles. In Contrastive Hebbian Learning, weight updates take place in positive (Hebbian) and negative (anti-Hebbian) phases. The weight update on the connection joining units i and j following the presentation of a training pattern is

$$\Delta w_{ij} \propto \breve{a}_i^+ \cdot \breve{a}_j^+ \cdot \Phi_{ij}(\breve{\phi}_i^+ - \breve{\phi}_j^+) - \breve{a}_i^- \cdot \breve{a}_j^- \cdot \Phi_{ij}(\breve{\phi}_i^- - \breve{\phi}_j^-), \tag{3}$$

where $\breve{}$ over a symbol refers to that quantity when the network has stabilized.

5 RESULTS

The following experiments were done on a network with a Visual Buffer of size 3x3. The Language layer contains four relation units for *on*, *under*, *left*, and *right*. Three dimensions were used to manipulate the difficulty of learning the different relations. Patterns for ON and UNDER are defined by three correlations: 1) *OM1* always correlates with the trajector, you can think of this as *size*; 2) *OM2* always correlates with ON and *OM3* always correlates with UNDER,these could be thought of as *movability* and *visibility* respectively; 3) Visually, they are identical.

OM2 and OM3 represent the fact that ON and UNDER are construed as different situations.OM1 and VB represent what ON and textscunder have in common.

The correlations for LEFT and RIGHT are: 1) At least one modality correlates with the trajector; 2) There are no situation-specific correlations; 3) Visually, they are identical.

1) means that there is no perfect predictor of which of the two objects in the scene is the trajector. 2) means that there is no perceptual reason to separate LEFT situations from RIGHT situations.

The point of this experiment is to show how learning the *left* and *right* is harder than learning *on* and *under*. The network is trained for two epochs on all the possible patterns allowed by the restrictions explained above using the Language layer as output.

The task is to, given a scene, describe it. A pattern is clamped in the non-linguistic layers of the network and the network is allowed to settle. The network succeeds when it turns on the RU that corresponds to the correct word in the Language layer.

For both ON and UNDER, it succeeds 100% of the time. For LEFT and RIGHT, it succeeds 55% of the time, 16% of the time it responds nothing and 27% of the time it gives the wrong word. This results show that *on* and *under* are easier to learn than *left* and *right*.

6 CONCLUSIONS AND FUTURE WORK

We have presented a model which explains the relative difficulty of learning spatial relation terms based solely on correlations between different sensory domains. By taking an "easy" (ON and UNDER) and "hard" (LEFT and RIGHT) relations, we show how the different correlation patterns affect the learning performance of the network. The network itself is relatively simple; it is able to mimic one aspect of the child's learning of spatial because of its built-in pattern of connectivity and because it includes explicit relation units.

The experiments we have conducted so far constitute simple demonstrations that the model is consistent. Next we will investigate the model's capacity to scale up to larger objects.

This will also allow us to test the extent to which the model generalizes to objects which it has not been trained on.

Our model makes predictions regarding the order of acquisition of spatial relations and also the sorts of relations which are lexicalized in the world's languages. We intend to explore these predictions as we continue to study ON, UNDER, NEXT TO, LEFT, and RIGHT and extend the model to other simple relations, including IN and BEHIND.

References

Behl-Chadha, G. & Eimas, P. (1995). Infant categorization of left-right spatial relations. *British Journal of Developmental Psychology*, *13*, 69–79.

Braine, L. . F. C. (1988). Context effects in left-right shape discrimination. *Developmental Psychology*, *24*(2), 183–189.

Clark, H. (1973). Space, time, semantics, and the child. In T. Moore (Eds), *Cognitive development and the acquisition of language*. New York: Academic Press.

Edelman, G. M. (1987). *Neural darwinism*. New York: Basic Books.

Fenson, L.and Dale, P., Reznick, J., Thal, D., Bates, E., Hartung, J., Pethick, S., & Reilly, J. (1993). *The macarthur communicative development inventories: user's guide and technical manual*. San Diego: Singular Publishing Group.

Gasser, M. & Colunga-Leal, E. (1997). Playpen: Toward an architecture for modeling the development of spatial cognition. Technical report, Indiana University, Cognitive Science Program, Bloomington, IN.

Gibson, E. (1969). *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts.

Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480–517.

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.

Logan, G. (1995). Linguistic and conceptual control of visual spatial attention. *Cognitive Psychology*, *28*, 103–174.

Movellan, J. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In D. Touretzky, J. Elman, T. Sejnowski, & G. H. on (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*. San Mateo, CA: Morgan Kaufmann.

Quinn, P., Cummins, M., Kase, J., Martin, E., & Weissman, S. (1996). Development of categorical representations for above and below spatial relations in 3- to 7-month old infants. *Developmental Psychology, 32*(5), 942–950.

Rudel, R.G. & Teuber, H. (1963). Discrimination of direction of line in children. *Journal of Comparative and Physiological Psychology*, *56*, 892–898.

Shastri, L. & Ajjanagadde, V. (1993). From simple associations so systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, *16*, 417–494.