Submitted to Cognitive Linguistics.

A Model of the Human Capacity for Categorizing Spatial Relations

Terry Regier
University of California at Berkeley

Abstract

Languages vary dramatically in their structuring of space. Despite this wide variation, however, the search for universals in spatial semantics is well motivated by the fact that all linguistic spatial systems are based on human experience of space, which is in turn constrained by the nature of the human perceptual system. I present a connectionist model which contributes to the search for universals in this domain. Its design incorporates a number of structural devices motivated by neurobiological and psychophysical evidence concerning the human visual system; these provide a universal perceptual core which constrains the process of semantic acquisition. Using these structures, the model learns the perceptually grounded semantics for closed-class spatial terms from a range of languages — providing at least a preliminary model of the human capacity for categorizing spatial events and relations. The model gives rise to two predictions concerning the manner in which one can expect to find motion encoded in closed-class spatial terms in the world's languages.

1 The Linguistic Categorization of Space

The linguistic categorization of space is a topic that has captured the attention of linguists and other cognitive scientists for a number of years. There is good reason for this. Spatial location is often expressed by closed-class forms, which have "the fundamental role of acting as an organizing structure for further conceptual material" (Talmy 1983:4). Thus, the use of these forms to denote spatial location is an indication that space has a privileged position as a foundational ontological category in language, a position which most other domains do not share. This point is strengthened by the fact that space often serves as a source domain for metaphoric understandings of other parts of the conceptual system; its influence is therefore not localized to an isolated sphere of experience (Lakoff 1987). In addition, while physical space is objectively measurable, human conceptualizations of space as manifested in language afford a good deal of subtlety in their semantic structure. In this respect space resembles the domain of color, another objectively measurable yet conceptually rich semantic domain (Berlin & Kay 1969).

Space is also attractive as a domain of semantic inquiry because there is considerable cross-linguistic variation in spatial systems. The notion of cross-linguistic variation in conceptual systems and modes of thought has long been a subject of fascination for linguists. Whorf (1956) is an early and extremely influential example of this. As we shall see, such differences in spatial systems are sometimes quite dramatic, but more often than not they are rather subtle, particularly when one compares closely related languages.

One language exhibiting a spatial system which is profoundly different from that of English is Mixtec, an Oto-Manguean language spoken in the state of Oaxaca, Mexico. Brugman (1983) presents a semantic analysis of spatial terms in Mixtec, spelling out the manner in which the body-part system is metaphorically mapped onto the spatial system, such that body-part terms are used

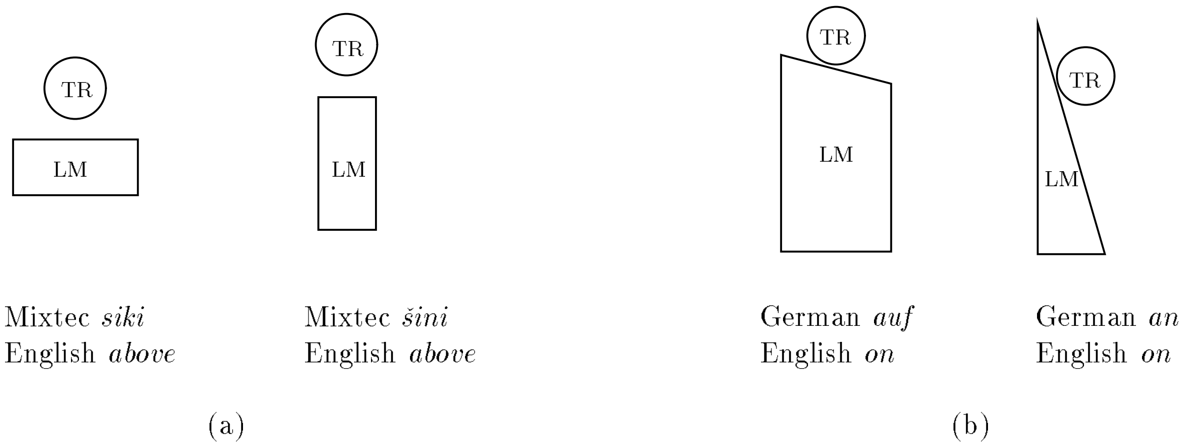


Figure 1: Examples from Mixtec and German

extensively as locatives. For example, to describe the location of a stone under a table in Mixtec, one would say

yuù	wã	híyaà	čìì-mesá
stone	the	be-located	belly-table

which, if translated word for word, would yield "The stone is located at the table's belly." In this sentence, Mixtec is schematizing the space underneath the table as the area adjacent to the belly of a quadruped standing on all fours, i.e. the region underneath it. The idea of metaphorical conceptualization of objects as animate beings, and the corresponding use of body-part names for regions of space near parts of objects, is by no means unique to Mixtec. In English *in back of* we have an instance of the same phenomenon. However, the metaphor is carried far further in Mixtec than in English or related languages. There is in fact no other way in Mixtec to express spatial relations.

It may seem at first that the Mixtec spatial system involves knowledge of the basic anatomy of quadrupeds and bipeds. However, it seems to be the case that much if not all of the categorization is based on fairly primitive perceptual features of the landmark object in question, such as the orientation of its major axis. Thus, a trajector above a long, wide landmark is considered to be located at the landmark's "animal-back" (Mixtec *siki*), by analogy to the dorsum of a horizontally-extended quadruped. By contrast, a trajector above a tall, erect landmark is considered to be located at the landmark's "head" (Mixtec *šini*), even if the landmark has no actual head. This distinction is illustrated in Figure 1(a). Note that both scenes would be classified as *above*, or *over*, in English.

More subtle examples of cross-linguistic variation in spatial systems can be seen when examining closely related languages (Bowerman 1989). Figure 1(b) shows two scenes which would both be classified as *on* in English, but which do not fall in the same category in German. In German, the orientation of the landmark surface which supports the trajector is significant, while it is not for English *on*. If the supporting surface is roughly horizontal, as in the scene on the left in Figure 1(b), the preposition *auf* is used to describe the relation. However, if the supporting surface is roughly vertical, as in the scene on the right, the preposition *an* is used instead (Stolcke 1990a).

A brief sampling of other languages which have been analyzed and which differ significantly from English in their structurings of space includes: Dutch (Bowerman 1989), Russian, Atsugewi (Talmy 1983), and Cora (Casad 1982). An exhaustive cataloging of cross-linguistic variation in spatial systems is well beyond the scope of this article.

Despite this variation, however, it is also extremely likely that commonalities exist across languages by virtue of the fact that all linguistic spatial systems are based on human experience of space, which is in turn constrained by the nature of the human perceptual system and the nature of the world around us. The point of interest here is that the varying spatial systems all derive from the same neural mechanisms and the same experiences with objects, gravity, and the like.

The domain of space thus suggests itself as an arena for explorations of issues of linguistic universality and variation with a force that few other domains can match: we know both that cross-linguistic variation exists, and that the essential sameness of human spatial experience across cultures motivates the search for semantic universals here. At the same time, since space is a fundamental ontological category and serves as a source domain for metaphorical mappings to many other domains in language, we are assured that inquiries concerning universality and variation in this domain will focus on elements that deeply affect the language as a whole, rather than just space itself.

2 Negative Evidence

As we have seen, the acquisition of linguistic spatial systems is of interest in its own right, because of the fundamental ontological status of such systems, and because of issues of universality and variation that arise in this domain. In addition to this, however, any account of the acquisition of such a system will also have to confront a very general issue in language acquisition, the problem of learning in the absence of explicit negative evidence.

Researchers in child language acquisition have often observed that the child learns language apparently without the benefit of explicit negative evidence (Braine 1971; Bowerman 1983; Pinker 1989). This introduces the following problem: if the child is never told that a particular utterance is ungrammatical, how does he or she learn not to utter it, while still learning to produce grammatical sentences that have also never been heard? How will he or she know which of these sentences that have never been heard conform to the constraints of the language being learned, and which do not? In other words, how does the child know not to undergeneralize or overgeneralize from the utterances heard, if nothing has been explicitly ruled out?

While the abovementioned researchers have focused on the "no negative evidence" problem as it relates to the acquisition of grammar, the problem is a general one, and appears in several aspects of language acquisition, including the acquisition of lexical semantics, with which we are concerned here. It is probably safe to assume that children are rarely, if ever, told that a particular configuration is not a good example of some spatial term, and yet they eventually learn to use these terms in novel situations without overgeneralizing or undergeneralizing. The problem is to determine just how this learning could take place. Any cognitive model of the acquisition of lexical semantics will have to come to grips with this issue, in addition to the issues of cross-linguistic variability touched on above.

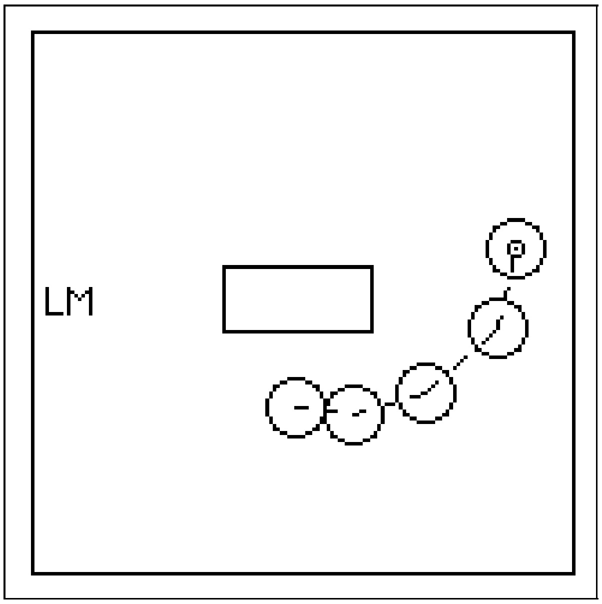


Figure 2: A movie: Russian *iz-pod*

3 Modeling the Human Capacity for Spatial Categorization

There are three central questions which serve as the primary motivating impetus for the computational work presented here: (a) What sort of system could adapt itself to the different structurings of space manifested in the world's languages? (b) How could such a system learn without explicit negative evidence? And (c) what could a model of this system tell us about semantic universals in the spatial domain? We shall be examining a connectionist model which provides answers to these questions.

Imagine a set of movies of simple 2-dimensional objects moving relative to one another, such that each movie has been correctly labeled as a positive instance of some closed-class spatial term from a particular language. The movies may be of arbitrary length. For example, Figure 2 presents such a movie, a positive example of the Russian preposition *iz-pod*, which has no single-word English counterpart, but translates to "out from underneath". The connectionist model presented here takes a set of such movies, each labeled as a positive example of some spatial term from some language, and learns the association between words and the events or relations which they describe. Once the system has successfully accomplished this task, it is able to determine which of the spatial terms learned would be appropriate for describing previously unseen movies. The system's task as a whole, then, is learning how to perceive simple spatial relations, both static and dynamic, so as to name them as a speaker of a particular language would.

The movie shown in Figure 2 was one of many used in training the system described here. Each movie contains a static object here referred to as the *landmark*, or *LM* (Langacker 1987); this is the reference object with respect to which other objects are located. In this movie, it is the horizontally extended rectangle in the middle of the scene. Each movie also contains another object, referred to as the *trajector*, or *TR*; this is the object located relative to the landmark. In this movie, the trajector is a circle moving from the region beneath the landmark, to the right and upwards. The dashed lines connect successive positions of the trajector as it moves. This particular movie is five

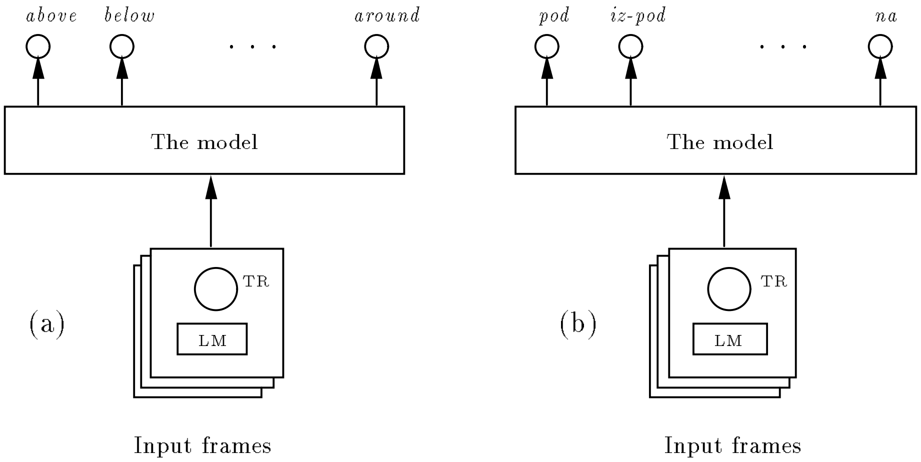


Figure 3: Model configurations for English and Russian

frames long, and here the frames are shown superimposed one on top of another. The final frame of the movie is indicated by a small circle located inside the trajector.

Figure 3 presents the model, configured to learn a set of English spatial terms in (a), and a set of Russian spatial terms in (b). The input to the model is a movie of the sort shown in Figure 2, and the model is trained so that when a movie portraying some event is shown, only those output nodes corresponding to closed-class forms which accurately describe the event are activated. For example, if the movie in Figure 2 were supplied to the model shown in Figure 3(b) after training, the *iz-pod* output node would become activated, indicating that the model has classified the movie as a positive example of *iz-pod*.

Clearly, if this system is to be taken as a model of the human capacity for linguistically categorizing spatial relations and events, it must be able to perform this learning task for spatial terms from any language. Furthermore, it should be able to learn without the benefit of explicit negative instances, as children appear to acquire language under those conditions. The model as currently constructed has learned systems of spatial terms from a range of languages with disparate categorizations of space, and has in some cases done so using positive evidence only. The primary interest of the model then is that it answers the three questions posed at the beginning of this section: (a) it provides an indication of what sort of system could adapt itself to the various structurings of space found in the world's languages, (b) it is able to learn using only positive evidence, and (c) as we shall see, at the same time it gives rise to predictions concerning semantic universals in the domain of spatial events and relations.

4 Incorporating Non-Linguistic Structure into the Model

The key to the model's design is its incorporation of a number of structural devices which are motivated by neurobiological and psychophysical evidence concerning the human visual system. This incorporation of non-linguistically motivated structures is very much in keeping with the spirit of inquiry into issues of semantic universality and relativity that motivates the work as a whole. After all, the central tension is between cross-linguistic variation on the one hand, and on the other, the knowledge that all languages share general human perceptual constraints on semantics, and are therefore very likely to have some semantic structure in common. The model presented here fits in as an element of this line of inquiry, as it is a linguistic learning system whose architecture is inspired by neurobiological and psychophysical insights into the workings of the human visual system. Thus, while the model can learn to adapt itself to any of a number of different structurings of space, as a language-learning human being does, there is an unchanging core of structure underlying the learning mechanism, corresponding to that perceptual structure which is common across all humans.

One critical point here concerns the status of the model: the use of perceptual structures such as the ones adopted here should not be taken to imply that this is a fully neural model, i.e. a complete reduction of linguistic processing to the neural level. It is rather a linguistic model which is merely motivated in its structure by disparate non-linguistic sources of evidence, some of them neurobiological in origin. This use of non-linguistic structures in a linguistic model is very much in the spirit of cognitive linguistics generally. From the point of view of cognitive linguistics, language is "inextricably bound up with psychological phenomena that are not specifically linguistic in character" (Langacker 1987:12). In the case of the model, it is very clear exactly which non-linguistic psychological elements are involved: those neurobiologically and psychophysically motivated visual structures which have been adopted as part of the overall design.

5 The Model

Figure 4 presents the model's architecture, with the non-linguistic structures we have been discussing highlighted. As shown in the figure, the model is configured to learn a set of English spatial terms. The model is trained using the connectionist error back-propagation algorithm (Rumelhart *et al.* 1986), while its architecture is informed by the design philosophy of *adaptive structured connectionism* (Regier & Feldman 1994). The fundamental idea behind this philosophy is the combination of elaborate motivated structure with the more traditional connectionist feature of adaptability to training data. The general idea of structuring connectionist networks is not new; see for example the work of LeCun (1989), Keeler *et al.* (1991), Mozer *et al.* (1991), and Guyon *et al.* (1991), among others. What distinguishes adaptive structured connectionism is its emphasis on the use of highly domain-specific and domain-motivated structure, rather than general techniques such as weight-sharing across temporal or spatial locations. The model illustrated in Figure 4 is an example of this: the neurobiologically and psychophysically motivated structures mentioned above are embedded in a trainable network whose parameters are adjusted during training to come to reflect the spatial structuring of a particular language.

Individual movie frames are presented to the system one by one, as shown at the bottom of Figure 4. At each time step, the hidden layer buffer labeled *Current* contains some learned representation of the current input frame, composed from features detected by the structured subnetwork below it. Above that, the structural devices shown in bold outline and marked (a), the *motion buffers*, serve to integrate information over the movie as a whole. At the end of the movie, those output nodes corresponding to spatial terms which accurately describe the event portrayed in the movie should be fully activated. Thus, if the movie was of a trajector moving over and across a landmark from left to right, then at the final time step, the nodes for *right* and *over* should be activated.

I shall be describing the motion buffers in some detail below, but shall first touch very briefly

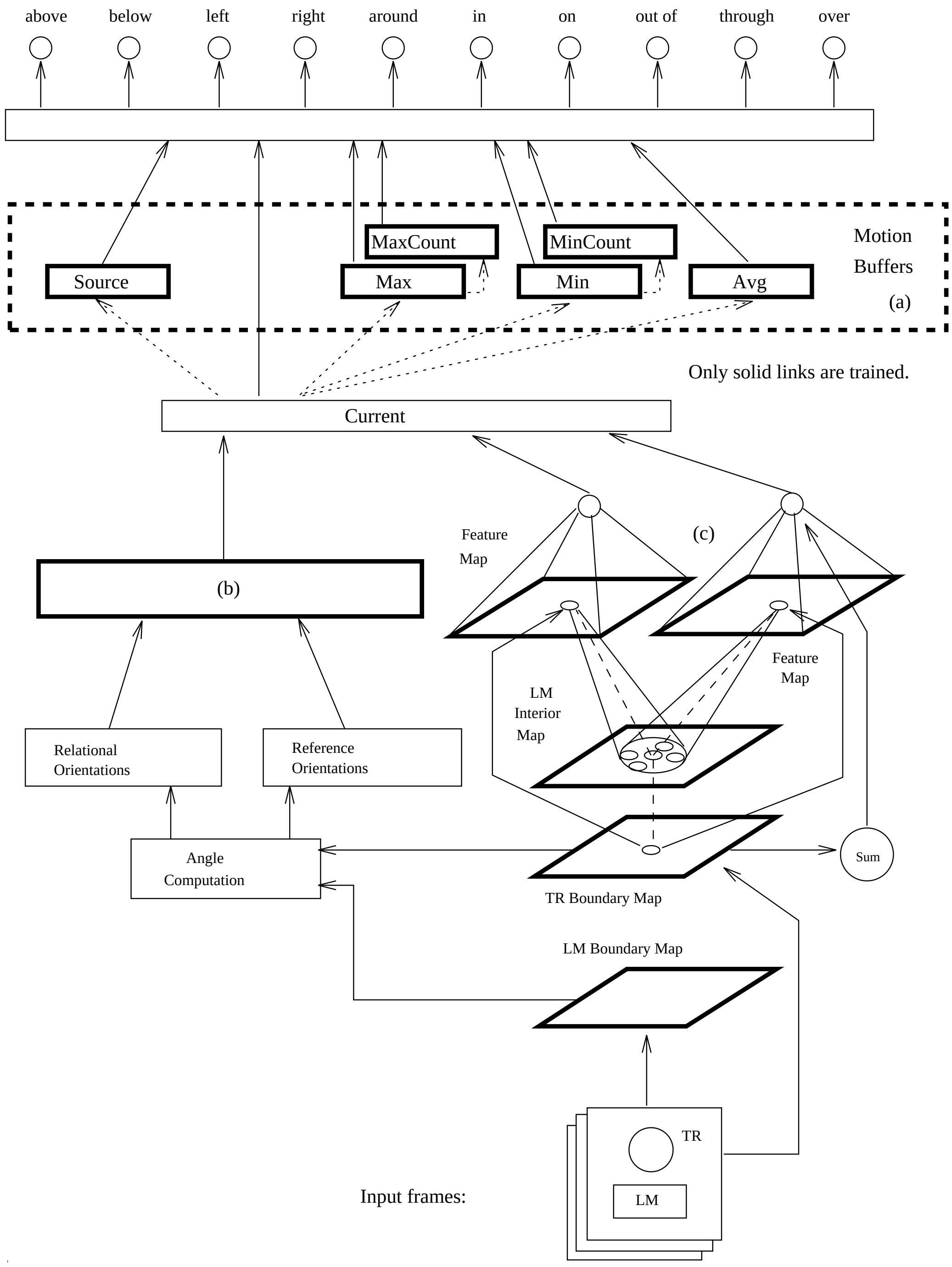


Figure 4: The model's architecture, with non-linguistic structures highlighted.

on the other motivated structures. The layer marked (b) in Figure 4 contains a set of *orientation sensitive cells*, i.e. cells which respond maximally when presented with a stimulus at a particular orientation. These are used for detecting the overall orientation of the trajector with respect to the landmark. The presence of orientation sensitive cells in primary and secondary visual cortex is well-established (Hubel & Wiesel 1959; Hubel & Wiesel 1962; von der Heydt *et al.* 1984).

The two topographic feature maps on the right of Figure 4, labeled (c), contain units with *circularly symmetric center-surround receptive fields*, i.e. units which respond maximally to particular circularly symmetric patterns in localized subregions of the topographic maps below. There is solid evidence for the existence of topographic maps in visual cortex (DeValois & DeValois 1990), and for the existence of cells with circularly symmetric center-surround receptive fields (Kuffler 1953). These feature maps are used in the detection of non-orientational spatial features such as contact and inclusion. The interested reader may find further details regarding these structures in (Regier 1992).

5.1 Motion Buffers: Tripartite Trajectory Representation

While the evidence for orientation sensitive cells, topographic maps, and cells with circularly symmetric center-surround receptive fields comes from neurobiology, the motion buffers are motivated by psychophysical and linguistic evidence. I present first the basic intuition which the motion buffers capture, then the evidence for structures of this sort, and then describe the function of these structures in some detail.

The motion buffers provide a tripartite trajectory representation for the model, such that the representation of a trajector moving relative to a landmark is composed of three subrepresentations: (a) a representation of the static relation between trajector and landmark at the *beginning* of the trajectory, i.e. the initial configuration of the two objects, (b) a representation of the static relation between trajector and landmark at the *end* of the trajectory, i.e. the resulting configuration of the two objects, and (c) a temporally integrated representation of the path from the beginning to the end of the trajectory. Thus, this can be viewed as a *beginning-path-end* representation, as it implicitly decomposes a trajectory into a beginning, an end, and the path from one to the other.

Psychophysical motivation for this structure comes from the visual phenomenon of apparent motion (Kolers 1972). Human subjects, when presented with an object displayed briefly at one position in the visual field, and then with another copy of the object displayed at another position, often perceive the object moving smoothly from the first point to the second. In the terms we are using here, given the beginning point and then the end point, the subjects perceptually infer the path. Thus, beginning point and end point serve as anchor-points of sorts, from which the path is computed. Recent connectionist research in modeling apparent motion (Olson 1989) has focused on the computation of explicit trajectories given beginning and end points.

There is linguistic motivation for this sort of structure as well: Lakoff (1987) and others have noted that linguistically, events are often structured in terms of three components: a *source*, or point of origin for motion, a *path*, or trajectory traced through space by the motion, and a *destination*, the endpoint of motion. The point here is that closed-class forms which refer to motion-based events are often easily characterizable in terms of these three elements. For our purposes here, the psychophysical motivation is more significant since we are primarily interested in independent non-linguistic motivation for the structures incorporated into the model, but the motivation is in fact dual in nature. One attractive although as yet unsubstantiated interpretation is to view these two sources of motivation as springing from a single underlying perceptual tendency to parse events into source, path, and destination.

This three-part structuring is reflected in the model's architecture. In particular, the *Source* buffer shown in Figure 4 will contain a representation of the starting configuration; at the last time step of the event the *Current* buffer will contain a representation of the ending configuration; and the remaining motion buffers (*Max, Min, Avg*, etc.) build up a static representation of all that has occurred over the course of the path. As we shall see, the nature of this static representation gives rise to constraints on what the model is capable of learning.

Recall that at each time step, the *Current* buffer will contain, by virtue of the adaptive structured subnetwork below it, a representation of the static features present in the current frame. So, to represent the configuration at start-event, all we need is a copy of what the *Current* buffer contained at the first time step of the movie. The buffer labeled *Source* in Figure 4 contains exactly that, and remains unchanged throughout the course of the movie. The link connecting the *Current* buffer to the *Source* buffer is shown as a dotted line, indicating that this link is not trained; rather, it performs a simple copy operation, and only on the first time step. Representing the configuration at end-event is even simpler: at the last time step of the movie, the contents of the *Current* buffer itself will be a representation of the configuration of the trajector relative to the landmark at the end of the event.

The remaining problem is representing the path, or trajectory, which the trajector traces out as it moves. This is done by the path buffers, the buffers labeled *Max, Min, MaxCount, MinCount,* and *Avg*. Note that the links leading into these buffers are shown as dotted lines, indicating that they are not trained.

The basic assumption behind the design of these path buffers is that it will suffice to keep track of what events have occurred over the path as a whole, without recording exactly when they occurred. In English *through*, for example, it is critical that the trajector be inside the landmark at some point during the path, i.e. after the source and before the destination, but it is not at all critical exactly where along the path. In order to record events in this manner, we track each of the learned features represented in the *Current* buffer, and record, for each such feature, the minimum, maximum, and average activations attained by the unit representing that feature over the course of the movie, together with the number of times the maximum or minimum value has been attained. There is a one-to-one correspondence between units in the *Current* buffer and units in each of these path buffers, such that a given unit in one of the path buffers will compute some function of the values seen in the corresponding *Current* buffer unit over the course of a movie. In particular, units in the *Max* path buffer record the maximum value attained by the corresponding *Current* buffer unit over the course of the movie. Similarly, units in *Min* buffer record the minimum value attained, and units in the *Avg* buffer record the average value. Units in the *MaxCount* buffer record the number of times the maximum value of a particular feature was attained, and the *MinCount* buffer operates analogously.

To make this somewhat more concrete, consider Figure 5. This figure illustrates the operation of the *Max* path buffer, which receives its input from the representation of the current input frame in the *Current* buffer. Since each unit of the *Max* buffer records the maximum value ever attained by the corresponding unit in the *Current* buffer, at the end of the event the *Max* buffer will contain a representation of the maximum value attained by each of the learned features which comprise the representation of individual input frames. The *Min* and *Avg* buffers similarly record the minimum and average values, respectively, attained by *Current* buffer nodes over the course of a movie.

Representing the trajectory in this manner has some serious ramifications. In particular, *the sequential order of events internal to the path is lost*: all we have available to us is the starting configuration, the ending configuration, and a non-sequential static representation of all that oc-

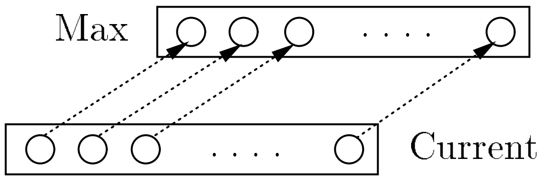


Figure 5: The *Max* buffer

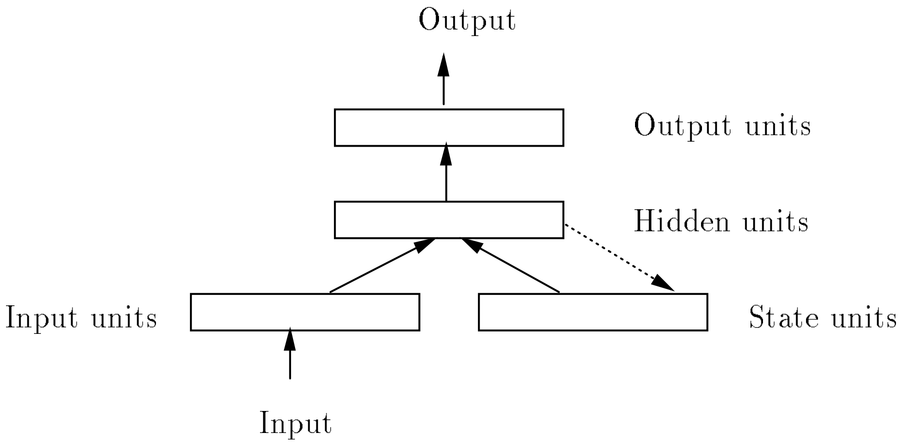


Figure 6: A simple recurrent network

curred over the path. I am postulating that this is sufficient for closed-class categorization of spatial events.

A general point to be made here is that different cognitive tasks often require different architectures. This point can be illustrated by comparing the motion buffer approach presented here with simple recurrent networks of the sort proposed by Elman (1988), and illustrated in Figure 6. These networks and others which are similar in design have gained considerable currency in connectionist sequence processing applications generally, despite their being appropriate only for a limited range of tasks. Simple recurrent networks copy their hidden unit representations down to a set of state units, which then supply input to the hidden layer at the next time step. This makes them similar in overall form to finite-state automata: at a given point in time, the output and next state depend on the current input and current state. Given their formal similarity to finite-state automata, it is not surprising that networks of this sort have had some success in the induction of regular grammars (Servan-Schreiber *et al.* 1988), and in the domain of grammatical induction generally (Stolcke 1990b). However, they are quite inappropriate for the task on which we are currently focusing, for at least two reasons. In the first place, there is no separate representation of the initial input, or the source in the terminology we are adopting here, so the network must learn to retain this information in its state representation whenever this is relevant, which it often is. This can add considerably to the difficulty of the learning task. In addition, in simple recurrent networks there is no straightforward representation of the fact that an event has occurred independent of the time step during which it occurred. These predicted disadvantages are borne out by empirical testing, which has shown that the motion buffer approach outperforms simple recurrent networks when applied to the task of learning closed-class categorizations of spatial relations (Regier 1992). Of course, the motion buffer approach, with its deliberate disregard for intermediate sequentiality of events, would be essentially useless for grammatical induction or other tasks in which temporally intermediate sequentiality is of the essence.

5.2 Learning without Explicit Negative Evidence

As noted above, any cognitive model of the acquisition of lexical semantics will have to provide some explanation of the manner in which children learn without explicit negative evidence. How could this sort of learning take place? One solution to the "no negative evidence" problem which suggests itself is to take every *positive* instance for one spatial term to be an *implicit negative* instance for all other spatial terms being learned. Thus, a positive example of *in* is taken as an implicit negative example of *above*, *below*, *outside*, *through*, etc. This is essentially the same as the "principle of mutual exclusivity", posited for the process of learning to name objects (Markman 1987). Related ideas can also be found in the language learning literature (Johnston & Slobin 1979; Clark 1987; Pinker 1989; MacWhinney 1989; Sinha *et al.* 1993).

There is a serious problem with this useful heuristic, however: it can easily give rise to *false implicit negative evidence*. For instance, a trajector can be both *above* and *outside* a landmark. So mutual exclusivity will be incorrect in this case, as it will take the positive example of *above* as an implicit negative example of *outside*. In fact, every positive example of *above, below, left, right,* and *on* will provide *false* implicit negative evidence for *outside*, under the mutual exclusivity heuristic.

This problem of false implicit negatives then is the central flaw in mutual exclusivity. Mutual exclusivity can be salvaged however, by treating positive and negative instances differently during training. In particular, implicit negatives are viewed as supplying only *weak* negative evidence, while explicit positives supply strong positive evidence (Regier 1992). This is the method adopted here in modeling the acquisition of spatial semantics. As we saw in Figures Figure 3 and Figure 4, the model learns a *system* of spatial terms in consort, so that a positive example of one spatial term can serve as weak implicit negative evidence for all others.

6 Results

The model was trained using back-propagation (Rumelhart *et al.* 1986) and quickprop (Fahlman 1988), a variant thereof. It has been tested on spatial terms from English, German, Bengali, Russian, and Mixtec. I present results from English, Russian, and Mixtec, to demonstrate that the model is capable of learning spatial terms from languages with widely varying spatial systems. The English spatial terms were learned from positive evidence only. In the cases of Russian and Mixtec, however, explicit negative evidence was also used since the unavailability of a full contrast set precluded the use of mutual exclusivity as a learning heuristic.

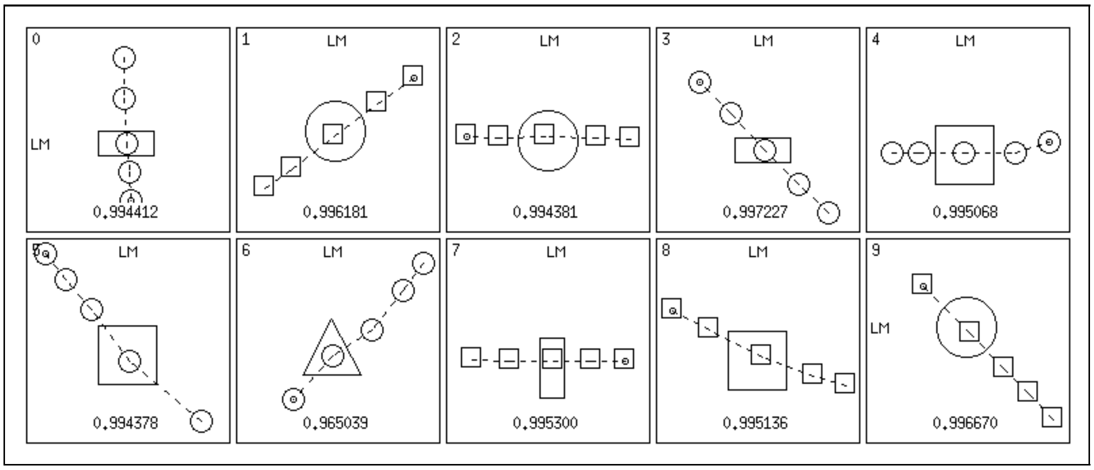


Figure 7: Positive examples of English *through*: a test set

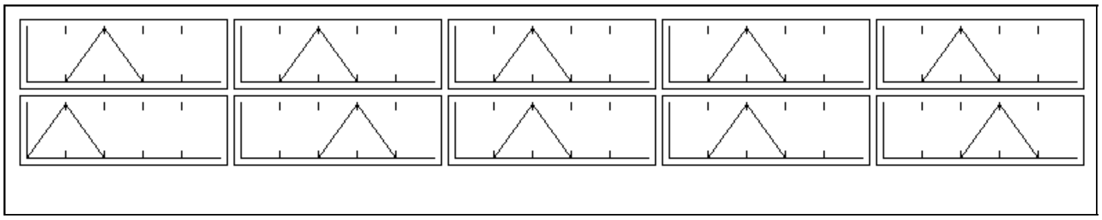


Figure 8: Response over time to *in* while viewing positive examples of *through*

6.1 English *through*

The model was trained simultaneously on the ten English spatial terms shown in Figure 4: *above, below, left, right, around, in, on, out of, through,* and *over* (in the sense of *over and across*). Learning took place without the benefit of explicit negative evidence.

Figure 7 presents the model's performance on a test set of positive examples of *through.* The number at the bottom of each movie is an indication of how good an example of *through* the model considers the movie to be. This ranges from 0.0 (poor) to 1.0 (excellent); the numbers indicate that the model has learned to give high ratings to good examples of *through.* In Figure 8, each box shows the model's response over time at the output node for *in* while viewing the positive example of *through* shown in the corresponding box in Figure 7. Each tick on the horizontal axis of boxes in Figure 8 represents a single time step, while the vertical axis in these boxes represents the strength of activation (0 to 1) of the output node for *in.* By comparing Figure 7 and Figure 8 one can see that the output node for *in* is strongly activated whenever the movie up to and including the current frame is a good example of *in*, particularly in its "motion-into" sense, as in *He walked in the room.* Figure 9 presents the model's performance on a test set of negative examples of *through*, indicating that the model consistently gives extremely low ratings for poor examples of *through*, as well as extremely high ratings for good examples. Learning was of essentially the same quality for the other spatial terms in the training set.

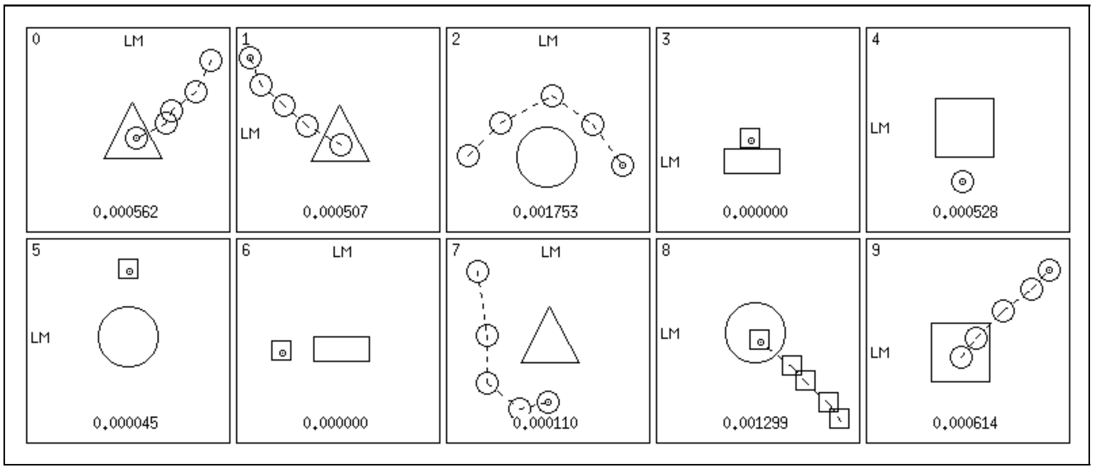


Figure 9: Negative examples of English *through*: a test set

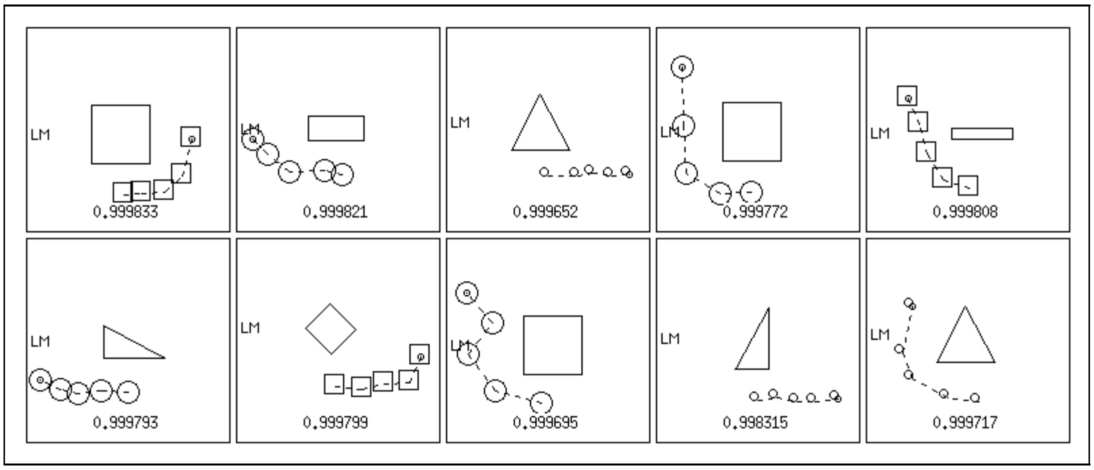


Figure 10: Positive examples of Russian *iz-pod*: a test set

6.2 Russian *iz-pod*

One respect in which the Russian linguistic categorization of space differs from that of English is that there is a single Russian word, *iz-pod*, which expresses "out from underneath", as was illustrated in Figure 2. The model was trained and tested on *iz-pod*, which it learned together with *pod* ("under"). Explicit negative evidence was used in this case. Figure 10 and Figure 11 present the model's judgments of how good an example of *iz-pod* each of the movies shown is. As these movies are from a previously unseen test set, this demonstrates that the model has learned to determine whether or not an event can be accurately described by *iz-pod*.

6.3 Mixtec *šini*

Finally, I present some static examples from Mixtec, so as to demonstrate the model's ability to learn a spatial system which is dramatically different from that of English. Recall that Mixtec

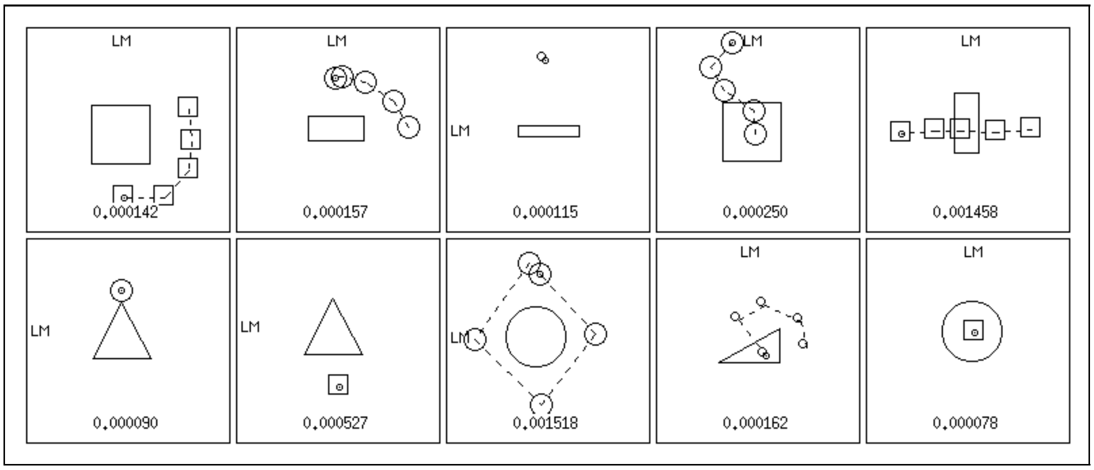


Figure 11: Negative examples of Russian *iz-pod*: a test set

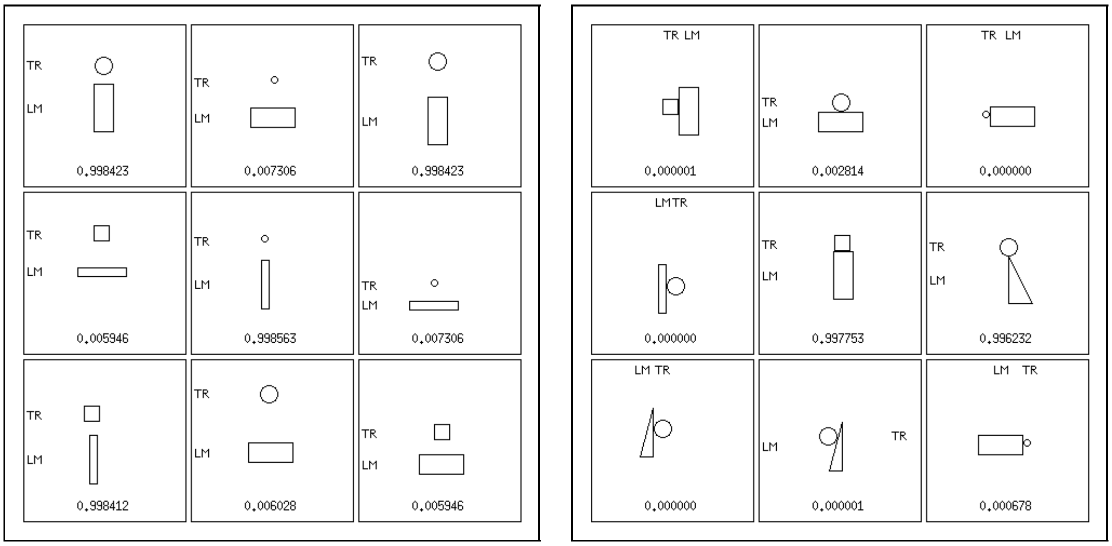
šini can be used to describe the relation between a vertically-extended landmark and a trajector either above it or on top of it. The model was trained on static *šini* alone, using explicit negative evidence, and then tested on scenes it had not seen previously, positive examples of English *above* and *on*.

Figure 12 presents the results of training, and highlights the differences between the Mixtec and English spatial systems. In (a) we see positive examples of English *above*; the numbers indicate how good an example of Mixtec *šini* each scene is. Similarly, (b) shows positive examples of English *on*, with an indication of how good an example of *šini* each one is. Clearly, Mixtec is sensitive to the orientation of the major axis of the landmark, as only those scenes with vertically upright landmarks are classified as *šini*. In adddition, notice that while English is sensitive to contact between landmark and trajector, Mixtec is not: if the landmark is vertically extended, then scenes in which the trajector is either above or on top of (and touching) the landmark will be considered good examples of *šini*.

7 Discussion

We have seen that the model is able to learn substantively varying spatial systems, in one case without the benefit of explicit negative evidence. While it would be premature to claim on the basis of these preliminary results that the model as currently constituted will be able to learn the spatial system of any human language, it can be taken as an initial model of the human capacity for closed-class expression of spatial relations.

It is useful to specify two different ways in which this model might fail, when faced with a new linguistic spatial system which it cannot learn. *Shallow failure* occurs when it becomes clear that it is necessary to add some structure to the model, but that structure fits in perfectly with the basic principles that guided the model's design. For example, if it were found that there were too few units in some subpart of the model as it stands, and that more of the same were needed, that would be an example of shallow failure, as the fix does not involve leaving the basic paradigm. *Deep failure*, on the other hand, occurs when it becomes evident that the fundamental approach to the model was wrong all along. For instance, if it were shown that tripartite trajectory representations



(a) (b)

Figure 12: Mixtec *šini* tested on positive examples of English *above* and *on*

---

or orientation sensitive cells were simply not appropriate at all, that would count as deep failure. While I have essentially no hope that the model as it is will avoid failure altogether, I do have hope that failure will be of the shallow variety. If this is the case, the work to date will still have provided some insight, by pointing out general principles that may well inform the human capacity for closed-class categorization of space.

A number of predictions fall out of this model. I shall focus on two, one of which results from the motivated structure which was built into the architecture, and one of which results from the model's structure together with the use of the mutual exclusivity heuristic. Thus, one of these can be seen as deriving purely from innate or universally acquired structure, while the other derives from universal structure and the use of a universally-posited learning heuristic.

7.1 The Intermediate Sequentiality Prediction

The *intermediate sequentiality prediction* is illustrated in Figure 13. It predicts that the only sequentiality which is relevant in closed-class forms for spatial events is the tripartite structure of beginning configuration, path taken, and ending configuration. In particular, no language will distinguish two events which differ only in the sequence in which configurations *within the path* occur.

For example, the model predicts that the two events in Figure 13 will not be distinguished on the basis of sequentiality by closed-class terms in any language. This is predicted since they are identical in their source-path-destination structure except for the order in which passing through the landmark and passing over a region of the landmark in mid-path occur. Note that beginning and end configurations are essentially identical in the two cases: in both cases, the trajector begins its path to the left of the landmark, and ends its path to the right of it. And in both cases, the path includes a segment during which the trajector is *inside* the landmark and one during which it is *above* the landmark. In fact, these two overall trajectories are distinguished only by the order

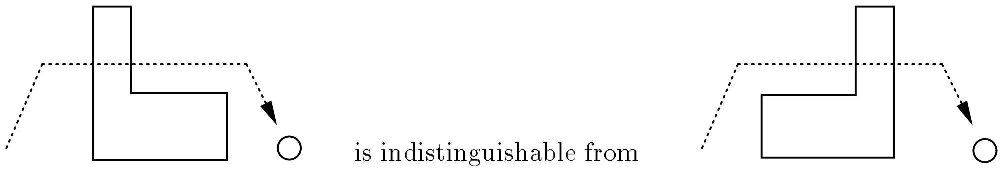


Figure 13: The Intermediate Sequentiality Prediction: Sequentiality within the path is irrelevant.

in which these path-internal *above* and *inside* segments occur: in the first case *inside* occurs before *above*, and in the second case the order is reversed. Since it is only path-internal sequentiality that distinguishes the source-path-destination structure of these two events, the prediction asserts that no language will distinguish these events on the basis of sequentiality using a closed-class form.

While this prediction is open to empirical falsification, it is interesting to note that there are languages with particularly rich systems of path specification which seem to conform to it. Atsugewi, for example, has a system of closed-class verbal suffixes indicating such intricate notions as "down into a gravitic container (e.g. a basket, a cupped hand, a pocket, a lake basin)", and "over the rim into a volume enclosure (e.g. a gopher hole, a mouth)" (Talmy 1983:15). These suffixes, although semantically quite elaborate, do not express path-internal sequentiality, but rather specify the nature of the landmark. Notice that even in the case of "over-the-rim into a volume enclosure" there is only one event specified as taking place during the path, namely going over the rim, and that intermediate sequentiality is therefore not relevant in this case.

One ramification of this prediction is that it implies that the description of complex paths in which path-internal sequentiality is significant will require the *composition* of closed-class forms, resulting in such constructions as "go over the river and then through the woods", where the individual closed-class forms *over* and *through* which serve as building blocks for the description as a whole do not express path-internal sequentiality, although the expression as a whole does.

This prediction falls directly out of the tripartite trajectory representation discussed above. Recall that although beginning-path-end sequentiality is preserved, all intermediate sequentiality is lost in the representation used here. All that is retained is a non-sequential representation of all that has occurred over the path as a whole. Thus the model will not be able to learn to distinguish events such as those shown in the figure. This is an example of a falsifiable prediction resulting from structures built into the model — the correlate in the model of either innate or universally acquired human perceptual structures.

7.2 The Endpoint Configuration Prediction

The endpoint configuration prediction, illustrated in Figure 14, states that any language which has a closed-class form denoting motion out of or motion through some configuration will also have a closed-class form denoting either motion into that configuration or static location in it.

Consider, for example, the English lexeme *through*, shown in the upper left of Figure 14. Here we see a small square trajector moving through a circular landmark object. We can view the lexeme *through* as denoting motion through a configuration of inclusion, in that inclusion of the trajector

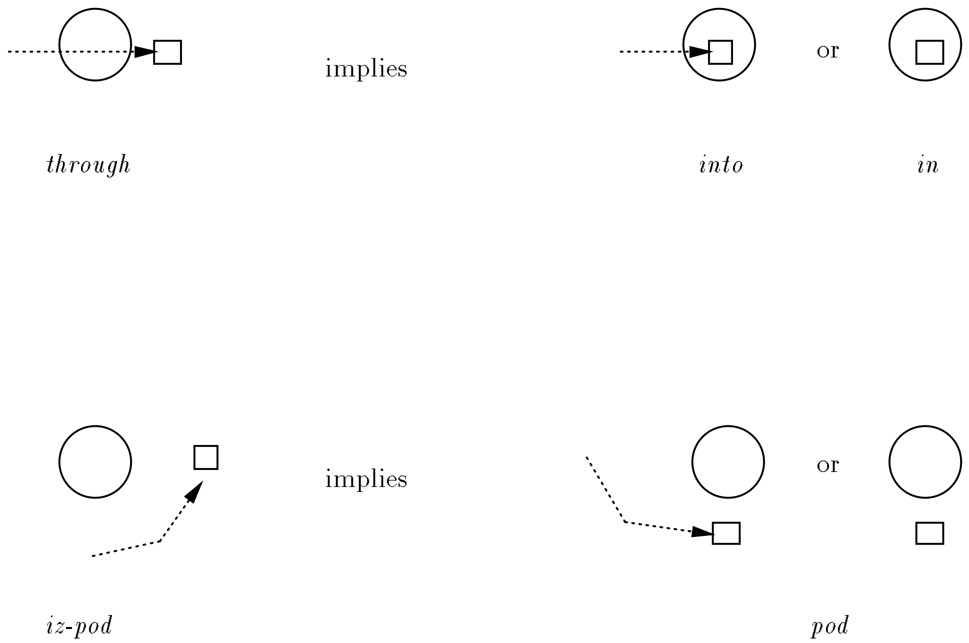


Figure 14: The Endpoint Configuration Prediction: Any language which has a lexeme denoting motion out of or motion through some configuration will also have a lexeme denoting either motion *into* that configuration or static location in it.

within the landmark occurs in mid-event for *through*. This leads us to predict, by our posited universal, that English will have a closed-class form denoting either motion into a configuration of inclusion (*into*), or static location in the configuration of inclusion (*in*).

Consider the lower example, from Russian. The fact that Russian has a closed-class lexeme *iz-pod*, denoting motion of a trajector out from underneath a landmark, leads us to predict that Russian will also have a closed-class form denoting either motion into the region under the landmark, or static location under the landmark. In fact Russian *pod* can be used in either of these two senses.

The model gives rise to this prediction because the attempt to build as structurally simple a model as possible resulted in a model with a particular constraint on its operation. Specifically, it resulted in a model that learns to categorize based on only those spatial features which occur at the end of some event it has seen. Thus, if it is never exposed to an event ending in the inclusion of the trajector in the landmark, it will not be able to make use of the static spatial feature of inclusion in the categorization of other events. This means it will not be able to learn *through*, as inclusion in mid-path is a critical element of *through*. And since the use of mutual exclusivity here means that the model is exposed to only positive instances of spatial terms from the language being learned, it will only see examples of events ending in inclusion if the language in fact has closed-class forms denoting such events. Thus, the existence of *through* in a language implies the existence of forms like *in* or *into*, as the model predicts that *through* would be unlearnable otherwise.

The essential idea behind this prediction, then, is that closed-class linguistic expression of motion out of or through some configuration is dependent upon the prior recognition that location in that configuration is linguistically relevant; in its turn, this recognition is dependent upon the existence in the language of terms denoting either location in that configuration or motion into it. The prediction suggests that terms denoting motion out of or through some configuration may be acquired after terms denoting location in or motion into that configuration, because of this dependency. The prediction is clearly falsifiable: any language containing a closed-class form for motion out of or through some configuration which did not also contain forms for motion into or location in that configuration would falsify it. And since it falls out of the perceptual endpoint emphasis of the model together with the principle of mutual exclusivity, it provides an illustration of determination of semantic structure by non-linguistic cognitive constraints and a universally-posited naming heuristic.

8 Conclusions

We have been concerned here with (a) what sort of system could adapt itself to the wide variety of structurings of space exhibited by human languages, (b) how such a system could learn without the benefit of negative evidence, and (c) what a model of this system could tell us about possible semantic universals. The connectionist model described here learns to linguistically classify perceived spatial events and relations, and does this for a range of languages with differing spatial systems. It also learns without the benefit of explicit negative evidence, as children appear to. It thus provides at least the beginning of an answer to the first and second questions.

The model answers the third question as well, in that it generates falsifiable linguistic predictions regarding semantic universals. The intermediate sequentiality prediction hypothesizes that the only sequentiality expressed by closed-class spatial forms is that of source-path-destination, and that all path-internal sequentiality is lost. This prediction arises from a universally-posited perceptual structure built into the model. The endpoint configuration prediction on the other hand hypothesizes that if a language has a closed-class form expressing motion out of or through some configuration, it will also have a closed-class form expressing motion into that configuration or static location in it. This prediction in its turn arises from an operating constraint on the model which resulted from the attempt to build as simple a model as possible, together with the use of the mutual exclusivity naming heuristic. Both predictions are left open to empirical falsification.

One interesting comparison with this work which suggests itself is the recent work of Landau & Jackendoff (1993). They analyze the English prepositional and nominal systems in terms of an underlying distinction in the visual system between neural structures which determine "what" an object is, and others which determine "where" it is. They associate the "what" system with object-denoting nouns, and the "where" system with locative prepositions. Clearly, their attempt to ground semantic structure in neurobiology is kindred in spirit to the approach taken here, although here, inspiration was drawn from biology (along with psychophysics and linguistics) for the benefit of what is at bottom a linguistic model, and no claim is made that the structural devices in this model correspond to identifiable neural structures. One of the central predictions which flows from Landau and Jackendoff's analysis is that only minimal representations of object shape are available to the closed-class locative system; this is similar in spirit to the intermediate sequentiality prediction here in that both predictions posit a glossing over of perceptual detail in the closed-class system. There are important divergences between the two analyses, however. Perhaps the most significant is that Landau and Jackendoff make universal claims based on an analysis of a single language, English, with only occasional references to non-English structurings of space, whereas this work has very deliberately taken a cross-linguistic approach. In addition, unlike theirs, this work is based on the analysis of a fully-specified computational model of the acquisition of closed-class spatial semantics.

The primary intention here has been to indicate that non-linguistic perceptual structures may affect linguistic semantic structure, through constraints on what is and is not learnable. This emphasis leads to a consonance in spirit with the philosophy of cognitive linguistics, which emphasizes the inter-relatedness of linguistic and non-linguistic phenomena. Interestingly, in its operation the model highlights a critical distinction between cognitive and generative linguistics. The object of study of generative linguistics is after all a characterization of the "innate component of the human mind that yields a particular language through interaction with presented experience, a device that converts experience into a system of knowledge attained: knowledge of one or another language" (Chomsky 1986:3). This is in fact precisely what the model presented here does in the domain of spatial semantics, but there is a crucial twist. While the model fits perfectly in the paradigm of postulating innate structures so as to enable learning by exposure to experience, and then searching for predicted universals, the innate structures posited here are not, as Chomsky envisioned them, linguistic principles. Rather, they are perceptual structures, independently motivated on neurobiological and psychophysical grounds.

9 Acknowledgements

Many thanks to the two reviewers for their comments and editorial suggestions, and to the members of the L-zero research group at the International Computer Science Institute in Berkeley for enjoyable and helpful conversations related to this work. Thanks to ICSI for office space and computational support.

References

Berlin, Brent, & Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution.* Berkeley: University of California Press.

Bowerman, Melissa. 1983. How do children avoid constructing an overly general grammar in the absence of feedback about what is not a sentence? Number 22 in Papers and Reports on Child Language Development, Stanford. Stanford University.

——. 1989. Learning a semantic system: What role do cognitive predispositions play? In *The Teachability of Language*, ed. by M. L. Rice et al, 133–169, Baltimore. Paul H. Brookes.

Braine, M. 1971. On two types of models of the internalization of grammars. In *The Ontogenesis of Grammar*, ed. by D. Slobin. Academic Press.

Brugman, Claudia, 1983. The use of body-part terms as locatives in chalcatongo mixtec. in Report No. 4 of the Survey of California and other Indian Languages, pp. 235-90. University of California, Berkeley.

Casad, Eugene, 1982. Cora locationals and structured imagery. Ph.D. dissertation, University of California, San Diego.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Clark, Eve. 1987. The principle of contrast: A constraint on language acquisition. In *Mechanisms of Language Acquisition*, ed. by B. MacWhinney, Hillsdale, NJ. Lawrence Erlbaum.

DeValois, Russell, & Karen DeValois. 1990. *Spatial Vision*. Oxford: Oxford University Press.

Elman, J. L. 1988. Finding structure in time. Technical Report 8801, Center for Research in Language, University of California, San Diego.

Fahlman, Scott. 1988. An empirical study of learning speed in back-propagation networks. Technical Report CMU-CS-88-162, Department of Computer Science, Carnegie Mellon University.

Guyon, I., P. Albrecht, Y. LeCun, J. Denker, & W. Hubbard. 1991. Design of a neural network character recognizer for a touch terminal. *Pattern Recognition* 24.105–119.

Hubel, D., & T. Wiesel. 1959. Receptive fields of single neurones in the cat's visual cortex. *Journal of Physiology* 148.574–591.

——, & ——. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* 160.106–154.

Johnston, Judith, & Dan Slobin. 1979. The development of locative expressions in english, italian, serbo-croatian and turkish. *Journal of Child Language* 6.529–545.

Keeler, James, David Rumelhart, & Wee-Kheng Leow. 1991. Integrated segmentation and recognition of hand-printed numerals. Technical Report ACT-NN-010-91, Microelectronics and Computer Technology Corporation.

Kolers, Paul A. 1972. *Aspects of Motion Perception*. Pergamon Press, New York.

KUFFLER, S. 1953. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology* 16.37–68.

LAKOFF, GEORGE. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

LANDAU, BARBARA, & RAY JACKENDOFF. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences* 16.217–265.

LANGACKER, RONALD. 1987. *Foundations of Cognitive Grammar I: Theoretical Prerequisites*. Stanford: Stanford University Press.

LECUN, YANN. 1989. Generalization and network design strategies. Technical Report CRG-TR-89-4, Connectionist Research Group, University of Toronto.

MACWHINNEY, BRIAN. 1989. Competition and lexical categorization. In *Linguistic Categorization*, number 61 in Current Issues in Linguistic Theory. John Benjamins Publishing Co., Amsterdam and Philadelphia.

MARKMAN, ELLEN M. 1987. How children constrain the possible meanings of words. In *Concepts and conceptual development: Ecological and intellectual factors in categorization*, ed. by Ulric Neisser, Cambridge. Cambridge University Press.

MOZER, MICHAEL, RICHARD ZEMEL, & MARLENE BEHRMANN. 1991. Learning to segment images using dynamic feature binding. Technical Report CU-CS-540-91, Dept. of Computer Science, University of Colorado at Boulder.

OLSON, THOMAS. 1989. An architectural model of visual motion understanding. Technical Report 305, Department of Computer Science, University of Rochester.

PINKER, STEVEN. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge: MIT Press.

REGIER, TERRY, 1992. *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization*. Computer Science Division, EECS Department, University of California at Berkeley dissertation. available as Technical Report TR-92-062, International Computer Science Institute, Berkeley.

——, & JEROME FELDMAN. 1994. Structured connectionist models and spatial concept learning. In *Handbook of Neuropsychology, Volume 9*, ed. by F. Boller & J. Grafman, 335–346, Amsterdam. Elsevier.

RUMELHART, DAVID E., GEOFFREY E. HINTON, & RONALD J. WILLIAMS. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, ed. by James L. McClelland & David E. Rumelhart, 318–362. MIT Press.

SERVAN-SCHREIBER, DAVID, AXEL CLEEREMANS, & JAMES MCCLELLAND. 1988. Encoding sequential structure in simple recurrent networks. Technical Report CMU-CS-88-183, Dept. of Computer Science, Carnegie Mellon University.

Sinha, Chris, Lis A. Thorseng, Mariko Hayashi, & Kim Plunkett, 1993. Comparative spatial semantics and language acquisition: Evidence from danish, english, and japanese. Revised version of paper presented at the International Conference on the Psychology of Language and Communication, Glasgow, September 1993.

Stolcke, Andreas, 1990a. (personal communication).

——. 1990b. Learning feature-based semantics with simple recurrent networks. Technical Report TR-90-015, International Computer Science Institute, Berkeley, CA.

Talmy, Leonard. 1983. How language structures space. In *Spatial Orientation: Theory, Research, and Application*, ed. by Herbert Pick & Linda Acredolo. New York: Plenum Press. also available as technical report 4, Institute of Cognitive Studies, University of California at Berkeley.

von der Heydt, R., E. Peterhans, & G. Baumgartner. 1984. Illusory contours and cortical neuron responses. *Science* 224.1260–1262.

Whorf, Benjamin Lee. 1956. *Language, Thought, and Reality*. Cambridge: MIT Press. (ed. John B. Carroll).