Grounded Acquisition of Containment Prepositions

Amitabha Mukerjee
Dept of Computer Sci and Engg
IIT Kanpur
amit@cse.iitk.ac.in

Mausoom Sarkar
Trilogy India
Pune
mausoom.sarkar@gmail.com

Abstract

We present a developmental approach towards a) pre-linguistic learning of grounded spatial schemas, and b) the acquisition of spatial prepositions based on association from these pre-linguistic concepts. We learn from sentential data, picking out the words most frequently associated with the concept, and show that simple associative structures are adequate for learning object names, or distinctions such as "in" or "out". A synthetic model of visual attention is used to constrain the set of objects in current focus. We first learn perceptual-object labels from simple 2D multi-agent visual streams co-occurring with word-separated utterances. We show that a notion of proximity between perceptual objects is sufficient to obtain a pre-verbal notion of graded spatial poses. We also demonstrate that the spatial concepts learned on a single shape generalize to other shapes, and correlate the learned perceptual schema with human precepts through a simple psychological test.

1 Introduction

The human infant acquires words from a grounded context, and forms perceptual schemas which reflect concepts that have arisen pre-linguistically, and are therefore independent of the words used. An important class of such concepts involves spatial relations, which are basic to acquiring motion verbs and other referents. The schema modeling such spatial relations are often fuzzy and although they may be approximated by discrete propositional structures, the grounded perceptual schema remains available as a fallback for disambiguating conflicts. Such models, often called *Image Schema* in Cognitive Linguistics (Langacker, 1999) or *Perceptual Schema* in Experimental Psychology (Mandler., 2004), involve abstractions on low-level features extracted from sensorimotor modalities. It is widely believed in cognitive science that the process of category formation operates on these inputs to define the structure of our conceptual space.

Today computational approaches to language are moving increasingly towards richer models of semantics, but hand-coded ontologies, like hand-coded grammars, appear to be inadequate to capture the richness of user experience. If we are to do with ontologies what we did with grammar, we may need to learn the semantics of linguistic tokens from grounded experience. In this process, it is important to start from the beginning, i.e. with the grounded acquisition of the first few hundred words, after which text processing (reading) can quickly accelerate this natural process. Without such grounded models, it is not clear how one can eventually construct the empirically validated ontologies underlying language.

How difficult is it to learn such semantics? In this work, we simplify the problem by providing a relatively simple grounding - a video well known in psychology (Heider and Simmel, 1944), from which the system clusters the spatial relations in an unsupervised manner based on a set of features which are specified. Now, when linguistic commentaries are added, the system finds that it is possible to first learn the symbol systems that map to the main objects (nouns). Next, we show that it is possible to learn mappings for basic prepositions like "in" and "on" related to containment.

The visual search is pruned using a computational model of visual attention. No knowledge of language or syntax is used at any point (except the fact that the input comes as word-separated chunks). Containment concepts are acquired as perceptual schemas specific to the particular shapes present in the input, but we find that

the results generalize well for other shapes, both for the containment object as well as the trajector. These schemas embody the relationship between the linguistic arguments (the containment object and the trajector), as well as the spatial relation (theme), hence it may also constitute a semantic basis for subsequently learning grammatical structures.

We consider the system to be at the stage of a learner who is able to identify frequent phoneme sequences from the speech stream as words.

1.1 Language Acquisition from Sentential input

Unlike other attempts which use single objects (Roy, 2000) or single words (Steels, 1997), we try to learn directly from sentential input which co-occur with a simple visual input containing multiple objects. Like (Ballard and Yu, 2003), our language learner learns from complete sentences (though not speech inputs), in a scene with multiple objects. Attentive focus is used to constrain the region of visual salience, and identify the constituents participating in an action. However, the learner is not in the presence of the speaker, and cannot follow cues from the gaze of the speaker to determine attentive focus. Instead, it is assumed that the learner realizes that her attentive mechanisms are similar to that of the speaker, a hypothesis we call the *Perceptual Theory of Mind*. The computational model of visual attention is based on the work of Koch and Ullman, as subsequently refined in (Itti, 2000). The model includes a parallel feature extraction stage, saliency map and winner-take-all (WTA) to obtain the conspicuous locations and an inhibition map which allows scanning of the maxima. Our gaze predictor incorporates extensions to dynamic image streams which adds motion features in the saliency computation, and confidence maps indicative of the positional uncertainty of visual objects.

Unlike the Ballard/Yu treatment of attentive focus to isolate pixels of the scene corresponding to a given term, our model uses motion based segmentation to isolate the perceptual objects. The problem of associating words with visual events is equivalent to a word correspondence problem in machine translation, as pointed out by (Duygulu et al., 2002), except that attentive focus permits much faster convergence even with the far simpler association measures than adopted there.

1.2 The role of Pre-Linguistic Concepts

The main distinction of this work, compared to other models of associative word learning (Roy, 2000; Regier, 1995) is that we assume that pre-linguistic notions of the concepts being acquired are already available. We first show how such concepts may arise from the perceptual stream alone. Thereafter, the language learner merely has to associate the concepts with their corresponding labels. The existence of pre-linguistic (perceptual) concepts for concrete objects is well known (Spelke, 1990; Bloom, 2000). Here we demonstrate that it is possible for a purely perceptual system to form notions of containment, well recognized as one of the earliest spatial concepts arising around the age of six months (Casasola et al., 2003).

Computationally, the first question we face is how to define a set of features for learning this notion. There have been many attempts at defining spatial relations based on different features. Excluding spatial formalisms that reduce an object to a point, we may consider the *Stolen voronoi area* approach (Edwards and Moulin, 1998), force histograms (R. Bondugula and Keller, 2004), or area-overlap features (Vorwerg et al., 1997). Of these, a cognitively plausible notion is that the learning agent has only a notion of proximity, based on which she can decompose the visual space into a influence regions corresponding to an area voronoi diagram. The introduction of a trajector into the scene can now be captured in terms of the Stolen Area (Edwards and Moulin, 1998) measure.

The basic developmental premise involving preverbal concepts, where categorical abstractions for spatial primitives are formed from pre-linguistic perceptual inputs, is a position that appears to be challenged by cognitive differences in spatial descriptors that arise between linguistic groups. For example, Korean speakers discriminate linguistically between tight- and loose-fit containment events, grouping tight-fit containment into the lexical category “kkita,” (along with tight-fit support events). English speakers do not perform as well as Korean speakers on non-linguistic categorization tasks involving such degree of fit relations (Bowerman and Choi, 2001). However, recent work seems to illustrate that pre-linguistic infants at the age of 5 months, from both English and Korean speaking backgrounds, are able to discriminate the degree-of-fit relations, but English learn-

ers appear to lose some of this discrimination by the time they start to speak (Hespos and Spelke, 2004). Thus there appears to be strong cognitive evidence that some spatial relations such as containment may be pre-linguistic.

1.3 Learning Spatial Prepositions

In his pioneering work on preposition grounding by (Regier, 1995), static and moving object scenes labeled with single words are used, and learning is achieved using a complex neural network based architecture inspired from neuropsychological and cognitive evidence. Orientation sensitive cells and centre surround maps calculated features used by motion buffers to segment the initial, final and in-between positions of the trajector. The various configurations and their corresponding closed class labels were learned.

In our model, the presence of prelinguistic clusters makes it unnecessary to learn the conceptual models from language, thus enabling far simpler algorithms running on much sparser data. Other account of prepositions (e.g. (Feist and Gentner, 2003), (Coventry, 1999)) propose to show how in/on judgments are affected by geometry and animacy in the figure and the function of the ground, and have led to the *SpaceCase* model (Lockwood et al., 2005). Here however, we are trying to simulate a far more primitive learner, and while prior concepts such as animacy may be available, functionality of figure/ground are often built based on the semantics of space, which is what we are trying to acquire here. We have therefore focused on simpler shape based measures.

As an unsupervised approach for discovering such pre-linguistic concepts, we use Self-Organizing Maps (SOM) (Kohonen, 1993). The emergent clusters are associated with the prepositions "in" and "out" from the running narrative using simple statistical measures based on mutual information. Both the pre-linguistic concepts as well as the verbal associations are then learned based on a single perceptual scene involving a single geometry and a few hundred configurations.

Another important issue for the beginning learner is the question of generalizing the notion of containment, learned on a single shape, to other shapes, and evolving it to eventually match the adult classification of containment. Here we show how such shape generalization capability may emerge and use five shapes BOX, DIAMOND, BOWL, MAZE, CSE-B, to test this generalizability.

The model assumes that the learner has the ability to segment the scene based on coherently moving blobs (Spelke, 1990), and that it has a measure of perceptual proximity. In addition, the learner has the ability to assign degrees of perceptual salience to different objects, which is simulated in this work by a computational model of dynamic visual attention.

Input: Heider and Simmel Video

We use a 2D video derived from the social psychology work of (Heider and Simmel, 1944). The co-occurring text were collected as part of an experiment on how users segment events into hierarchical subtasks (Martin and Tversky, 2003)[1]. In this task, users were asked to segment the actions in the scene and also to describe the action in an unconstrained narrative. Consequently, the linguistic input has the wide variety expected in multiple articulations for the same scene (see Table 1 below).

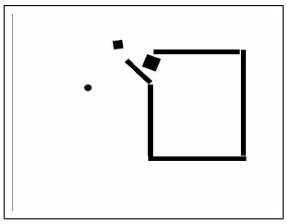

Figure 1: *Input Video.* Scene from the Chase sequence (derived from (Heider and Simmel, 1944) - recreated by Bridgette Martin (Martin and Tversky, 2003)). Three agents, "big square", "small square" and "circle" play and chase each other.

2 Synthetic models of Visual Attention

Computational models of Visual Attention involve bottom-up and top-down processes. While top-down processes vary depending on task requirements, bottom-up aspects are more stable and have been encoded for static images (Itti, 2000) based on parallel extraction of intensity, colour and orientation contrast feature maps. Colour and intensity contrast maps are obtained as feature pyramids (maps at different scales), along with center-surround maps (multi-scale difference of feature maps). The center-surround feature processing is

[1]We are grateful to Bridgette Martin Hard and Barbara Tversky for sharing this video as well as the Hide and Seek video, both of which were prepared by Bridgette and her colleagues, and also for the transcriptions of the co-articulated text collected by her.

Start Frame	End Frame	Subject One	Subject Two
617	635	the little square hit the big square	they're hitting each other
805	848	the big square hit the little square	and they keep hitting each other
852	1100	the big square hit the little square again; the little circle moves to the door; the big square threatens the little circle	now the circle is blocking the entrance for the big square; now the circle is inside the square
1145	1202	the big square goes inside the box; (and) the door closes	another square went inside the big square

Table 1: *Description of Events in the [Chase video].* Differing statements by two subjects.

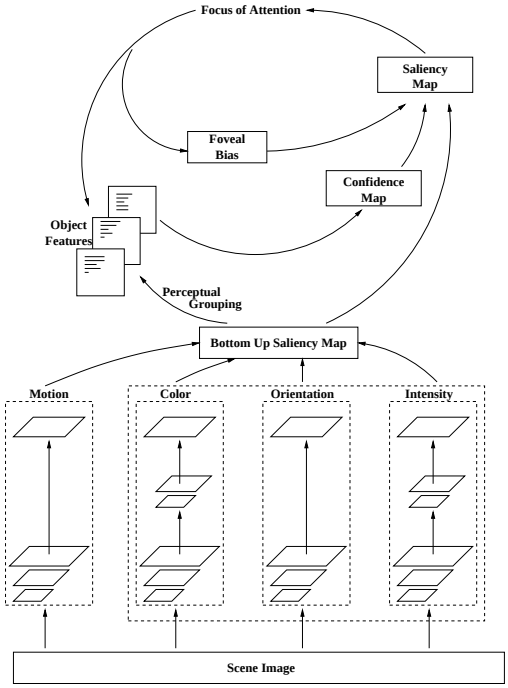

Figure 2: *Bottom-Up Dynamic Visual Attention Model*. The feature maps for static images (colour, intensity and orientation) are extended with a motion saliency map (based on optical flow). In addition a confidence map records which sites have not been visited for a longer time. Winner-Take-All determines the next fixation.

similar to the difference of gaussian convolved images (DOGs). For orientation specific processing, gabor filters are used with different frequencies and at different scales to generate the orientation specific feature map.

The static model, which replicates saliency map structures likely to be present in the LGN or V1 regions of the mammalian cortex, has been extended (Singh et al., 2006) to model dynamic scenes based on motion saliency. Motion saliency is computed from the optical flow, and a confidence map is introduced to record the uncertainty accumulating at scene locations not visited for some time. A small foveal bias is introduced to mediate in favour of proximal fixations as opposed to large saccadic motions. The saliency map is the sum of the feature maps and confidence maps, mediated by the foveal bias, and a Winner-Take-All (WTA) isolates the most conspicuous location for the next fixation. The overall architecture is shown in Figure [2]".

2.1 Perceptual Theory of Mind

The Theory of Mind hypothesis (Bloom, 2000) holds that the learner has a model for several aspect of the speaker's mind, at various levels from a sensitivity to the object being attending to, to belief structures (e.g. children under three are found to be incapable of entertaining false beliefs). In this work, we focus at the lowest end of this spectrum, and focus on what we call the *Perceptual Theory of Mind*. While much of the Theory of Mind work has focused on gaze following based on cues from the speaker's eyes or her gaze direction, the Perceptual Theory of Mind makes a much weaker claim: in the absence of direct cues from a speaker, it assumes that the speaker would have attended to those parts of the scene that the learner also finds salient. This is probably a valid assumption for children from the age of about one year 18 months onwards (Flavell, 2004), although the mechanisms for perceptual salience are themselves being developed at this stage. In our work, we do not specify a particular development status for our learning agent, but assume this model to infer that the scene objects being attended to by the agent were also salient for the speaker at the moment of utterance.

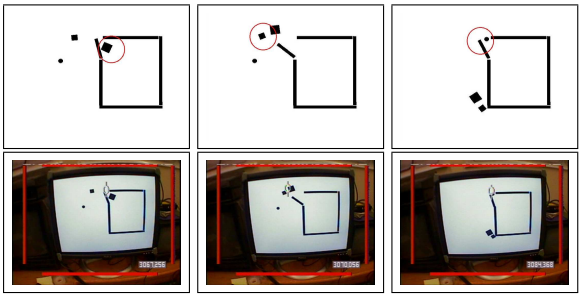

Figure 3: *Focus of attention* in three scenes as computed by the synthetic attention model (top row, circles) and as determined by gaze tracker on human viewer (bottom row, tall oval).

Language Acquisition experiments tend to cast doubts on the efficacy of a purely associationist model of learning words, and it is true that a large percentage of our vocabulary is not learned using multimodal inputs but from reading. Nonetheless, this work presents some evidence that for the beginning learner, multimodal associations mediated by attentional processes provide strong and reliable cues for learning nominals and their properties, verbs and their argument and event structures.

3 Learning Containment Descriptors

Before we learn spatial relation labels, we need to identify the nouns describing the participants in the relation. In considering the word association task, one may assume that the learner has been exposed to some other linguistic fragments, so that highly frequent words like "the" and "is", which appear in many other contexts, are known to be more general, and are not applied to this situation. (In the British National Corpus, "the" occurs 1500 times more frequently than "square"). Using perceptual equivalence relations based on shape, we associate the objects with words from a second video, *Hide and Seek* (another 2530 frames), using the probability measures outlined below.

The word-object association is estimated using the product of mutual information of word w_i and object o_j with their joint probability.

$$A = \Pr(w_i, o_j) \log \frac{\Pr(w_i, o_j)}{\Pr(w_i)\Pr(o_j)}$$

We calculate the product of joint probability and mutual information because if W and O ($W = \bigcup_i w_i$ and $O = \bigcup_i o_i$) are two random variables then their Mutual Information $I(W, O)$ would be

$$I(W, O) = \sum_i \sum_j \Pr(w, o_j) \log \frac{\Pr(w_i, o_j)}{\Pr(w_i)\Pr(o_j)}$$

and $\Pr(w_i, o_j) \log \frac{\Pr(w_i, o_j)}{\Pr(w_i)\Pr(o_j)}$ would be the contribution of each word object pair. Results show a high degree of correlation (Figure 4).

In the spatial domain, there have been several attempts at defining spatial relations involving continuum measures defined over different geometric features on object pairs. Many of these measures involve point attributes such as potential fields, but our interest here is more on area measures since perceptually the objects constitute an area.

The assumption of the existence a proximity notion required a spatial representation (data structure) which can represent proximity between objects. The voronoi model of space is one structure

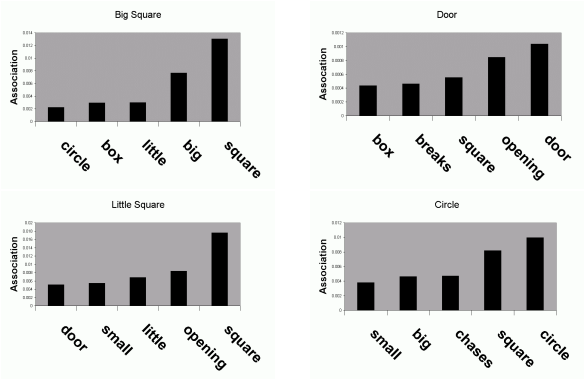

Figure 4: *Association of nominals with the visual objects.* Big Square, Door, Little Square and Circle.

which can store proximity information. It stores each site into a cell which is a locus of all points closer to the enclosed site. The voronoi boundary between two sites or objects shows adjacency between those objects and the degree of proximity can be found using voronoi influence zones or *Stolen area* (Edwards and Moulin, 1998; Edwards et al., 1996). Our feature set uses the voronoi model and is defined in terms of how much of the zone of influence is lost by the insertion of the trajector. The distinction between closed (bounded) zones and open (unbounded) ones constitute a key perceptual signature of containment.

The learned image schema for relative spatial poses is in the nature of a topographical neural net, also known as a kohonen map or a Self-Organizing Map (SOM). This is an unsupervised clustering method and requires no labels or other priors. Once the clustering is obtained, it can be mapped to linguistic data with very few exemplars.

In our work, we attempt to look at just two spatial terms related to the containment concept - the prepositions "in" and "out", which are among the earliest spatial terms learned (Bowerman and Choi, 2001).

3.1 Spatial Prepositions of Containment

Grounded learning of spatial modifiers involves defining concept classes to represent spatial distinctions. Clearly these classes are not discrete but graded (or continuum). As stated earlier, we chose the (stolen voronoi area) model of spatial relation (Edwards and Moulin, 1998; Edwards et al., 1996)as this requires very little of the learner other than the capacity to identify the most proximal region, and it can also give a graded measure of membership.

A voronoi representation of space divides the space into tessellations (voronoi regions) where each region contains a site - traditionally, this is a point, but it may also be a line or an area. The division of space is based on the notion of proximity such that all points in the voronoi region are closer to its site than to any other site. Voronoi diagrams in fig 5a, b shows a division of space in terms of points proximal to each of the point sites, similarly fig 5c is a division in terms of points proximal to either the point site or a line site. The voronoi diagram for any arbitrary shape can be determined by defining it as a voronoi diagram of points and line segments. In our model for line segments, we distinguish (as do most algorithms) the two end points from the body of the line; this results in two internal boundaries for the voronoi diagram - the separator between this end-point and the interior of the line - and are useful in discriminating the nature of the intersection between zones, and are retained.

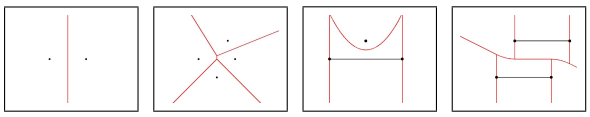

Figure 5: *Voronoi diagrams of points and line segments:* Voronoi diagrams of a,b) point sites c) a Point and a line segment d)two line segments. Line segments are bounded by two end "points", resulting in two internal boundaries (as in *c* and *d*)

The voronoi model of a space changes as agents move in it. This dynamic nature results in qualitative changes as the boundary between two sites or objects (reflecting adjacency of those objects) shifts to some other objects. For the *empty* space, each part of the boundary has its zone of influence. Some zones may be bounded, reflecting proximity with other objects, where as some zones may be unbounded, reflecting an open nature. As an agent enters this space, the areas of these zones are reduced or "stolen", and these reductions function as features of spatial pose (Edwards and Moulin, 1998), and these features are used to try to learn the containership concept.

3.2 Learning Containment

Containment is one of the earliest concepts in our repertory, yet it offers tremendous complexity. In adult usage, containment is affected by function which results in a wide variety of ramifications (Coventry, 1999). However, to an early learner, the prototypical interpretation of containment and container emerges based on abstracting on percepts, possibly earlier than six months (Casasola et al., 2003).

The voronoi model used here as the learning feature captures the influence of one object over another. We use the ratio of stolen area from bounded and unbounded voronoi regions as a binary feature. The ratio is calculated between the voronoi area consumed by the introduction of the new object and the initial voronoi area of the reference object fig 6, and distances between features are determined in terms of an euclidean metric to construct the kohonen map.

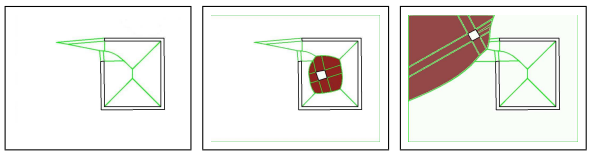

Figure 6: *Stolen Voronoi area*. Initial voronoi regions (a). Dark area in (b) and (c) shows the stolen area.

3.3 Self Organizing Map

A self organizing map is an unsupervised learning method based on associative memory, also known as the kohonen map (Kohonen, 1993). A grid of neurons is defined, each with a k dimensional weight vector during the learning process so that proximal regions represent "proximal" patterns in the input data. From an initial assignment of random weights, each data point finds its best matching unit/neuron (BMU) and the weights of neurons in the neighbourhood of BMU are modified towards the weight vector of the input, using the equation

$$W(t+1) = W(t) + \theta(t)\alpha(t)(D(t) - W(t))$$

where $W(t)$ is the weight vector of the neuron, $D(t)$ id the input vector, $\theta(t)$ is the neighbourhood function which constraints the amount of influence on the neighbours, and $\alpha(t)$ is the learning rate.

We use a simple kohonen map with a gaussian neighbourhood function and a linearly decreasing learning rate. The grid size of 30x40 nodes represents the clustering space. The SOM visualizations are shown using the U-matrix approach which colours each cell according to the the sum of the weights of all its neighbors. The 3D landscape is visualized by mapping the U-matrix

height into the RGB space. uniform patches of colour represent zones of uniform height reflecting the possibility of a cluster.

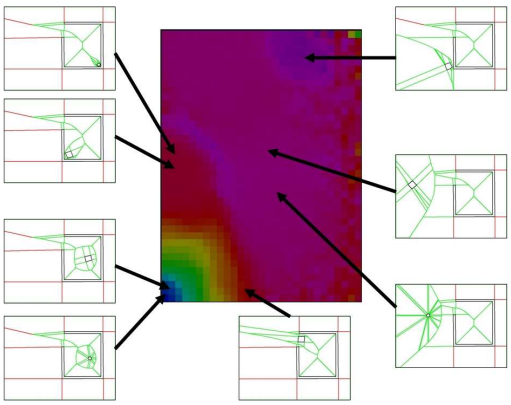

Figure 7: *SOM map of containment relations*. The Self-Organizing Map acts as our perceptual schema. Different configurations of the trajector (square) have different excitations in the SOM, reflected in a colourized map. The same map can identify a cluster for an input pose, or given a conceptual description it can generate a maximum likelihood pose.

Maps were trained based on input data of different spatial poses obtained from the video. Figure 7) shows a stabilized topographical distribution of nodes in the feature space. Different positions of the trajector w.r.t. the Enclosure(ENC) are shown in different parts of the SOM map. Among the regions that may be thought of as "in", the points at the center are furthest from the out region, and the points near the door (bottom-center configuration in fig.7) show a sharp gradation in colour. Similarly, the out regions are also graded by distance from the center. We also find finer distinctions emerging that may be taken as "corner" or "center". For our task, we cluster this space using k=2, and two clusters emerge, roughly corresponding to configurations inside and outside the ENC.

The "perceptual schema" represented by the SOM is learned only from a single trajector on a single ENC. Words in a language however, refer to a class and not such a specific instance. How general is the learned perceptual schema? Changes in trajector shape result in a slight redistribution of the stolen areas, and do generalize well, but changes in ENC are more complex. How do these schemas generalize to general shapes? We explore this question in section 4.

3.4 Mapping to Prepositions "in" and "out"

These spatial relation clusters are then matched with all words occurring in the user commentaries, using the same mutual information measure as used for nouns. No grammatical or other knowledge is assumed, but the nouns which have been already learned are kept out of the words considered for this association.

Since containment structures involve two objects, phrases containing two nouns/object names were considered and the feature vector relating the two objects was calculated and checked for the cluster they lie in. Similarly cluster labels for each phrase were collected and automatas were constructed to represent the transition between clusters for each phrase (e.g. circle coming out of the room will transition from one state (in) to another (out)). These automatas were then associated with their corresponding phrases. Unlike noun learning, very frequent words were not eliminated, since in (top 10) and out (top 50) are themselves very frequent.

Strong associations emerge between the words "out" and "in", and the state transitions "in to out" and "out to in" respectively (fig 8). It seems that while talking about a state transition, destination state has communicative saliency, and matches the matches the descriptive label. Words like "of" are seen to be in a tie with "out" - this may be indicative of the English construct "out of". However, the particle "of" occurs almost twice as frequently as "off" in general text, so it is likely to drop off in the specificity measure if we were to include other (non-spatial) contexts.

4 Context Sensitivity of Containment Schema

An important aspect of any concept and one that contributes significantly towards the brittleness of machine approaches is the degree of reliance on context. In this instance, the shape of the enclosing object ENC is an important and immediate aspect of the context. In the training data the enclosure was a squarish room with an opening the top left. We investigated the effect of changing the enclosure shape by computing the voronoi features on four different ENC shapes.

Visualization of the feature vector calculated for two different shapes shown in fig 9. The red area is towards the "in" region and "out" is indicated by a preference for blue.

Notice the colour gets darker near edges and corners because the amount of stolen voronoi area decreases.

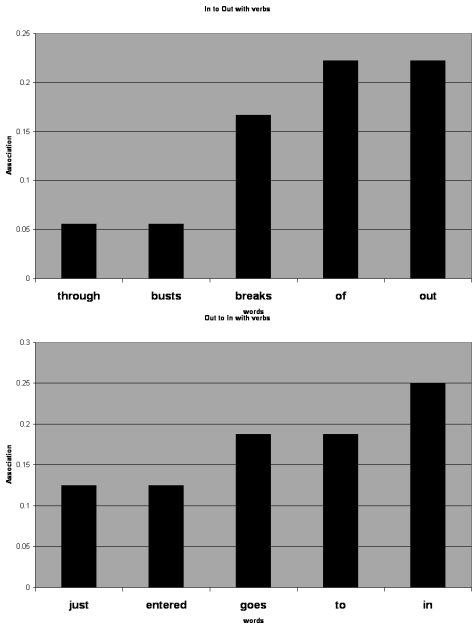

Figure 8: *Preposition Learning.* Matches between clusters and words in text category in-to-out (top); category out-to-in (bottom).

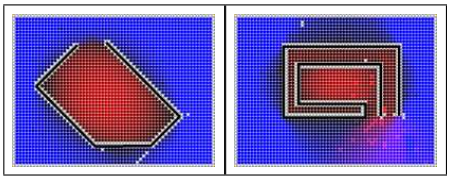

Figure 9: *Visualization of the feature vector.* a) DIAMOND shaped enclosure, and b) MAZE shape. The function does not generalize well to the maze; but then neither did humans in our psychological tests.

Classification of regions as "in or out" was done using the SOM. The results are shown in fig10. The feature vectors at each point were the inputs to the SOM and colour was assigned according to the top 5 best matching units (SOM neurons) and their cluster.

An artifact of the feature computation that may affect some results is the bounding box over which the computation is performed. Cognitively, it appears, that this bounding box is not infinite - i.e. each object exerts influence only to a certain distance. In fig 11 the shape has nearly parallel voronoi edges. If the bounding box is large enough (fig 11b), it would enclose the complete bounded area or else it may be taken as a part of unbounded area 11a.

Figure 12 shows the feature computation for a shape with some nonconvexities. We use bounding boxes which are 2x and 8x the size of the minimum enclosing box for the shape. With the smaller box, the lower area beneath the obtuse angle is clearly going to be "out"; in the larger box it becomes doubtful. The circular space at the top has a higher degree of "in"-ness in the first reading. This sort of confusion was also found in our human subjects (section 5).

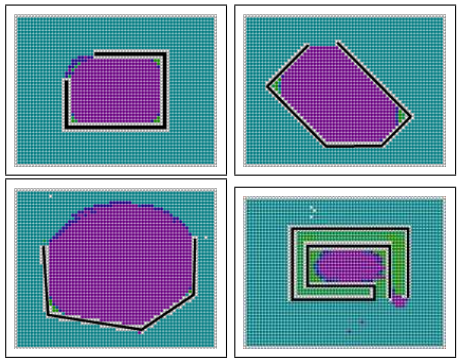

Figure 10: *SOM Perceptual Schema output.* a) Shape BOX, b) Shape DIAMOND, c) Shape BOWL, and d) Shape MAZE. The SOM is undecided in the maze, but considers the space in the inner cavity as "in".

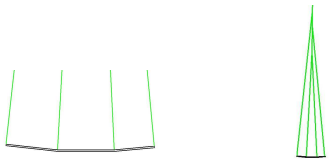

Figure 11: *Bounding box "Zone of Influence" effects.* Consider the three edges at the bottom. a) With a tight bounding box, the voronoi edges hit the boundary and appear to be unbounded; b) with a larger bounding box, the voronoi edges meet and constitute a bounded voronoi area.

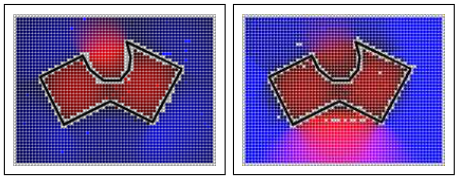

Figure 12: *Effect of bounding box on non-convex boundaries.* Here, the visualization is different for the Shape CSE-B (floor plan of building). The lower part gets be classified "in" for the map on the right (loose bounding box) because that area is bounded in a wide space. On the other hand, when the zone of influence is tight, the lower area is marked as "out". Some of this confusion also persists in human judgments.

5 Validation through Psychological Tests

In order to validate the semantic reality of the model learned, we compared the perceptual schema with human responses on the set of shapes. Twenty student volunteers in the age group 19-23 were shown images of various shapes with a circle at randomly generated positions. They were

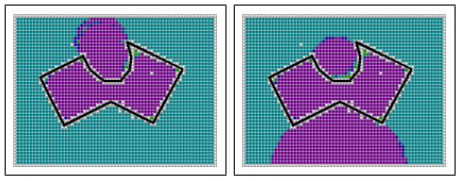

Figure 13: *SOM Perceptual Schema output for* CSE-B, a) SOM with bounding box = 2 $\times$ minimum enclosing box (tight); b) SOM with 8x box (loose).

asked to decide whether it was "in" or "out" of the box. For each configuration the "in"-ness was computed as the percentage of in responses. This was compared with the in/out ratio calculated by using the top 5 best matching units (SOM neurons). The comparative data is presented in in table 2, 3.

The results reveal that mismatches are higher with shapes like CSE-B (and also MAZE), which are the furthest from the convex box that was the basis of the learning. However, in verbal responses, humans also indicate less confident about their choices in these shapes than in the other (convex) shapes. The zone of influence effect can be seen at the top and the bottom of the CSE-B shape. For the MAZE, containment decision may require other features - based on path continuity, as opposed to merely proximity measures.

Shape	human	SOM (2x)	SOM (8x)
	90	100	100
	90	100	100
	90	100	100
	90	100	100
	35	100	100
	10	20	20
	0	0	0
	0	0	0
	0	0	0

Table 2: *Degree of "in"-ness.* Human responses vs SOM for Shape BOWL. Results from placing trajector in several random positions; human responses tabulated against the learned perceptual schema. Overall, a good degree of match.

Shape	human	SOM (2x)	SOM (8x)
	5	100	100
	100	100	100
	5	40	40
	5	40	20
	100	100	100
	0	0	100
	0	20	0

Table 3: *Degree of "in"-ness for Shape* CSE-B. human responses vary a good bit from the learned schema. There is considerable uncertainty among humans also regarding these containment for these complex shapes.

6 Conclusion

We have developed (from an extremely simple visual stimulus) a coherent approach to acquiring object labels and spatial prepositions that build on simple spatial features. In particular, we have made no prior assumptions on knowledge about the agents or domain of action; the only abilities inherent in the learner is that it has a model of visual attention, and that it can identify perceptually proximal regions. Based on this, we show how pre-verbal conceptual schema may arise, which are then mapped into labels for objects (nouns) and spatial poses (prepositions).

Another key notion explored in this work is a concretization of the image schema. Here it is stored in a topographical neural net, which has strong cognitive plausibility (Kohonen, 1993), and is also one of the more effective unsupervised clustering methods. A crucial aspect of the current work is that the pre-linguistic concepts were learned from a small corpus of data and could be generalized to many novel shapes (and object poses).

However, the perceptual schema learned are clearly not independent of context, and although some variation in context is tolerated (e.g. to other convex enclosures), the system will need to be retrained for more general instances. Instead of 81 seconds of learning, however, the human learner has days and months and years of exposure, and clearly this can leqad to the construction of extremely rich and diverse schemata. In the context of computational applications of language, such schema can be maintained much more easily than most traditional systems and provide a simple mechanism for updating world ontologies in an empirically validated manner.

Learning the semantics immediately identifies the participants in the prepositional predicate as two distinguished entities - a trajector and an enclosure or container. This leads to an asymmetric "in" relation between square and box, which is at the heart of the predicate in(circle, box). These predicates are more flexible and robust than hand-coded ones, and may thereby provide a beginning for bottom-up semantic processing for information retrieval, complementing present top-down approaches.

The predicate and its argument structure are also key notions in composing the notion of containment with other notions - and this process

gives rise to different structures in different natural languages in what is known as grammar. It may be possible to apply such schema towards compositional operations - based on argument sharing, conceptual blending, and other cognitive mechanisms, leading to a semantic set of constraints on experience, which may underlie the constraints on language that we call grammar (Langacker, 1999).

The processes for combining schema is a key hurdle in this path, and one that deserves close attention. Some pointers to grammar discovery (on a very different path) can be seen in the work of (Ford, 2003). In our case, once the verbal heads of various phrases are known, and with some knowledge of the semantics of closed-class words, it would be possible to identify some of the roles played by grammatical elements in different constructions appearing in a narrative.

Another extension is to consider other spatial modifiers such as directions (left, right), transitional modifiers (around, through), etc. Concepts referring to corner or center showed up in primary perceptual schema as peripheral and central regions - but when we tried to transfer this (esp. the corners) to other shapes, they did not generalize well. It would be important to explore the conditions under which generalized corner and edge concepts can be learned. Also work on verbs can be extended using the spatial concepts learned in this manner.

References

Dana H. Ballard and Chen Yu. 2003. A multimodal learning interface for word acquisition. In *International Conference on Acoustics,Speech and Signal Processing(ICASSP03)*, volume 5, pages 784–7.

Paul Bloom. 2000. *How Children Learn the Meaning of Words*. MIT Press, Cambridge, MA.

Melissa Bowerman and Soonja Choi. 2001. *Language acquisition and conceptual development*. Cambridge University Press.

Marianella Casasola, L.B.Cohen, and E.Chiarello. 2003. Six-month-old infants categorization of containment spatial relations. *Child Development*, 74:679–693.

Kenny R. Coventry. 1999. Function, geometry and spatial prepositions: Three experiments. *Spatial Cognition and Computation*, 1:145–154.

Pinar Duygulu, Kobus Barnard, J.F.G. deFreitas, and David A. Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV-02*, pages IV:97–112, London, UK. Springer-Verlag.

Geoffrey Edwards and Bernard Moulin. 1998. Towards the simulation of spatial mental images using the voronoi model. In P.Olivier and K-P. Gapp, editors, *In Representation and processing of spatial expressions*, pages 163–184. Lawrence Erlbaum Associates, Mahwah, NJ.

G. Edwards, G. Ligozat, A. Gryl, L. Fraczac, B. Moulin, and C. M. Gold. 1996. A voronoi-based pivot representation of spatial concepts and its application to route descriptions expressed in natural language. In *In Proceedings, 7th International Symposium on Spatial Data handling, Delft*, pages 7B1–7B15, The Netherlands.

Michele I. Feist and Dedre Gentner. 2003. Factors involved in the use of in and on. In *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society*.

J.H. Flavell. 2004. Theory-of-mind development: Retrospect and prospect. *Merrill-Palmer Quarterly*, 50:274–29.

Dominey Peter Ford. 2003. Learning grammatical constructions in a miniature language from narrated video events. In *cognitive science*.

F. Heider and M. Simmel. 1944. An experimental study of apparent behavior. In *American Journal of Psychology*, volume 57, pages 243–59.

Susan J. Hespos and Elizabeth S. Spelke. 2004. Conceptual precursors to language. *Nature*, 430:453–455, 22 JULY.

L. Itti. 2000. *Models of Bottom-Up and Top-Down Visual Attention*. Ph.D. thesis, California Institute of Technology, Pasadena, California.

Teuvo Kohonen. 1993. Physiolocigal interpretation of the self-organizing map algorithm. *Neural Networks*, 6:895–905.

Ronald W. Langacker. 1999. *Grammar and Conceptualization*. Mouton de Gruyer, Berlin/New York.

K. Lockwood, K. Forbus, and J. Usher. 2005. Spacecase: A model of spatial preposition use. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society.*, Stressa, Italy.

Jean Matter Mandler. 2004. *Foundations of Mind*. Oxford University Press.

Bridgette Martin and Barbara Tversky. 2003. Segmenting ambiguous events. In *Proceedings of the 25th annual meeting of the Cognitive Science Society*. Crucial for our Data-Collection chapter.

P. Matsakis R. Bondugula and J. Keller. 2004. Force histograms and neural networks for human-based spatial relationship generalization. In *Proceedings of Int. Conf. on Neural Networks and Computational Intelligence*.

Terry Regier. 1995. A model of the human capacity for categorizing spatial relationships. *Cognitive Linguistics*, pages 63–88.

Deb Roy. 2000. Integration of speech and vision using mutual information. In *International Conference on Acoustics,Speech and Signal Processing(ICASSP00)*.

Vivek Kumar Singh, Subhransu Maji, and Amitabha Mukerjee. 2006. Confidence based updation of motion conspicuity in dynamic scenes. In *Third Canadian Conference on Computer and Robot Vision CRV 2006*.

Elizabeth S. Spelke. 1990. Principles of object perception. *Cognitive Science*, 14:29–56.

Luc Steels. 1997. Language learning and language contact. In *Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks, ECML-97*, pages 11 – 24.

Constanze Vorwerg, Gudrun Socher, Thomas Fuhr, Gerhard Sagerer, and Gert Rickheit. 1997. Projective relations for 3d space: Computational model, application, and psychological evaluation. In *Proceedings of the AAAI*, pages 159–164.