Spatial Prepositions and Vague Quantifiers: Implementing the Functional Geometric Framework

Kenny R. Coventry[1], Angelo Cangelosi[2], Rohanna Rajapakse[2], Alison Bacon[1],
Stephen Newstead[1], Dan Joyce[3], and Lynn V. Richards[1]

[1] Centre for Thinking and Language, School of Psychology, University of Plymouth, Drake Circus, Plymouth, PL4 8AA, United Kingdom
{kcoventry, a1bacon, snewstead}@plymouth.ac.uk
[2] Adaptive Behaviour & Cognition Research Group, School of Computing Comms & Electronics, University of Plymouth. Drake Circus, Plymouth, PL4 8AA, United Kingdom
{acangelosi, rrajapakse}@plymouth.ac.uk
[3]Institute of Cognitive Neuroscience, University College London, Alexandra House, 17 Queen Square, WC1N 3AR, United Kingdom

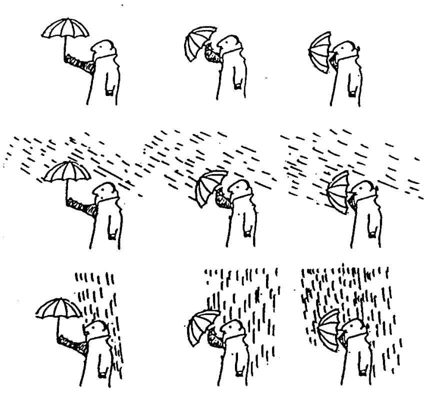
Abstract. There is much empirical evidence showing that factors other than the relative positions of objects in Euclidean space are important in the comprehension of a wide range of spatial prepositions in English and other languages. We first the overview the *functional geometric framework* (Coventry & Garrod, 2004) which puts "what" and "where" information together to underpin the situation specific meaning of spatial terms. We then outline an implementation of this framework. The computational model for the processing of visual scenes and the identification of the appropriate spatial preposition consists of three main modules: (1) Vision Processing, (2) Elman Network, (3) Dual-Route Network. Mirroring data from experiments with human participants, we show that the model is both able to predict what will happen to objects in a scene, and use these judgements to influence the appropriateness of *over/under/above/below* to describe where objects are located in the scene. Extensions of the model to other prepositions and quantifiers are discussed.

1 Introduction

Expressions involving spatial prepositions in English convey to a hearer where one object (located object) is located in relation to another object (reference object). For example, in *the coffee is in the cup*, the *coffee* is understood to be located with reference to the *cup* in the region denoted by the preposition *in*. Understanding the meaning of such terms is important as they are among the set of closed class terms which are generally regarded as having the role of acting as organizing structure for further conceptual material (Talmy, 1983). Furthermore, from the semantic point of view spatial prepositions have the virtue of relating in some way to visual scenes being described, and therefore measurable characteristics of the world (Regier, 1996). Hence, it should be possible to offer more precise semantic definitions of these as opposed to many other expressions because the definitions can be grounded in perceptual representations.

Most approaches to spatial prepositions have assumed that they only require coarse grained properties of the objects involved as constraints on their use (e.g., Herskovits, 1986; Landau and Jackendoff, 1993). Computational models too have made the same assumption, and have focused on mapping individual prepositions onto geometric computations in the scene being described (e.g., Logan & Sadler, 1996; Regier, 1996; Regier & Carlson, 2001; Gapp, 1995). Yet there is now much evidence (see Coventry & Garrod, 2004, for a comprehensive review) that "what" objects are influences how one talks about "where" they are. For example, Coventry, Prat-Sala and Richards (2001) found that acceptability ratings of sentences such as the *umbrella is over the man* were influenced by whether the objects in the scene were shown to be fulfilling their protection (or containment) functions. For instance, with reference to the scenes shown in Figure 1, sentences were rated as being significantly more appropriate when the umbrella was depicted as protecting the man from rain (scenes in the middle row), and least appropriate when the rain was falling on the man (scenes in the bottom row). Furthermore, extra-geometric variables came into play even when the prototypical geometric constraint for the use of a term holds (i.e., effects were found even for scenes in the first column). Additionally, Coventry et al. found that function has a much bigger affect on the ratings for *over/under* than for *above/below*, and conversely that geometry (e.g., rotation of the umbrella in Figure 1) influences the ratings of *above/below* more than *over/under*.

Fig. 1. Example scenes used by Coventry, Prat-Sala and Richards (2001)



Similar effects have been found across a wide range of prepositions and methodologies. For example, extra-geometric effects have been found for *in* and *on* (Coventry, Carmichael & Garrod, 1994; Coventry & Prat-Sala, 2001; Garrod, Ferrier & Campbell, 1999; Feist & Gentner, 1998), *above* (Carlson, Covey & Lattanzi, 1999; Carlson & Tang, 2000), *over* (Coventry & Mather, 2002*), in front of* and *behind* (Carlson-Radvansky & Radvansky, 1996), *between* (van der Zee, Watson & Fletcher, in press; Coventry & Garrod, in press), and *near* (Ferenz, 2000). Furthermore, the effect sizes found across these studies indicate that these effects are not minor pragmatic add-ons to geometric formulations, but rather indicate that extra-geometric variables are central to the comprehension and production of spatial terms.

1.1 The Functional Geometric Framework

Reviewing the extra-geometric evidence, Coventry and Garrod (2004, in press) classify these influences into two types; *dynamic-kinematic routines*, and *conceptual knowledge* regarding the specific functions associated with specific objects. Dynamic/kinematic routines implicate knowledge of what will happen to scenes over time, and the initiation of such routines is related to knowledge of *what* objects are in the scene. In particular these dynamic/kinematic routines relate to Jeannerod's (1994, 2001) distinction between "semantic" visual representations, usually associated with visual imagery, and "pragmatic" representations associated with motor imagery. Jeannerod assumes that motor images underlie such things as preparing for an action or rehearsing an action. Furthermore he argues that the two representations, the semantic and the pragmatic, have a neural correspondence with the *what* and the *where* systems described above. Whereas "semantic" representations encode relatively detailed information about objects in a scene, "pragmatic" representations encode visual properties in relation to affordances, i.e., those visual characteristics that are important in organizing motor programs for manipulating the objects. These include information about the size, weight and shape of objects, as well as special features of those objects that are relevant for their manipulation, such as the location of handles for grasping. Empirically Freyd, Pantzer and Cheng (1988; see also Schwartz, 1999) carried out experiments in which they observed systematic memory errors for scenes involving the same objects in the same geometric configurations, but with different forces acting on them. Thus, in a situation where a plant pot is first seen supported by a chain then not supported, observers tend to misjudge the position of the plant pot as being lower in a subsequent memory test. In the spatial language domain, Coventry (1998) and Garrod, Ferrier and Campbell (1999) have demonstrated similar effects for *in* and *on*. For example, using static scenes involving pictures of ping pong balls piled high in containers with a string attached to the top ping pong ball in many scenes, they found that ratings of the appropriateness of *in* to describe such scenes was directly correlated with independent judgments of the likelihood that the ball and container would remain in their same relative positions over time should the container be moved.

In addition, a great deal of specific knowledge about objects is also required. For example, the same convex object labelled a *dish* versus a *plate* is clearly associated with the expectation of a containment versus a support relation (Coventry, Carmichael & Garrod, 1994). Similarly, knowing that jugs are primarily containers of liquids has been shown to weaken *in* judgements for solids piled high in a jug as compared with *in* judgements for the same pile in a bowl with the same degree of concavity (Coventry, Carmichael & Garrod, 1994; Coventry & Prat-Sala, 2001).

Coventry and Garrod (2004) argue, importing terminology from Ullman (1984, 1996), that the application of geometric and dynamic-kinematic routines underlie the comprehension of spatial prepositions. Furthermore, the application of such visual routines is driven by knowledge of the objects involved in the scene and how those objects typically interact in past learned interactions between those objects. Furthermore, just as objects are associated with particular routines, both geometric and dynamic-kinematic, prepositions themselves have weightings for these parameters. As we have seen above, the comprehension of *over/under* is better predicted by extra-geometric relations than the comprehension of *above/below*, while conversely the comprehension of *above/below* is better predicted by geometric routines than the comprehension of *over/under*. In the functional geometric framework it is how these constraints "mesh" together (cf. Glenberg, 1997; Barsalou, 1999) that underpins the comprehension of spatial prepositions. The computational model we next outline implements the multiple constraint satisfaction in the functional geometric framework and maps onto new and existing datasets from human participants. The approach introduces cognitive-functional constraints by extennding Ullman's (1996) notion of visual routines to include operations on dynamic rather than static visual input. We next outline the components of the model, together with the experimental data used to test and validate the model.

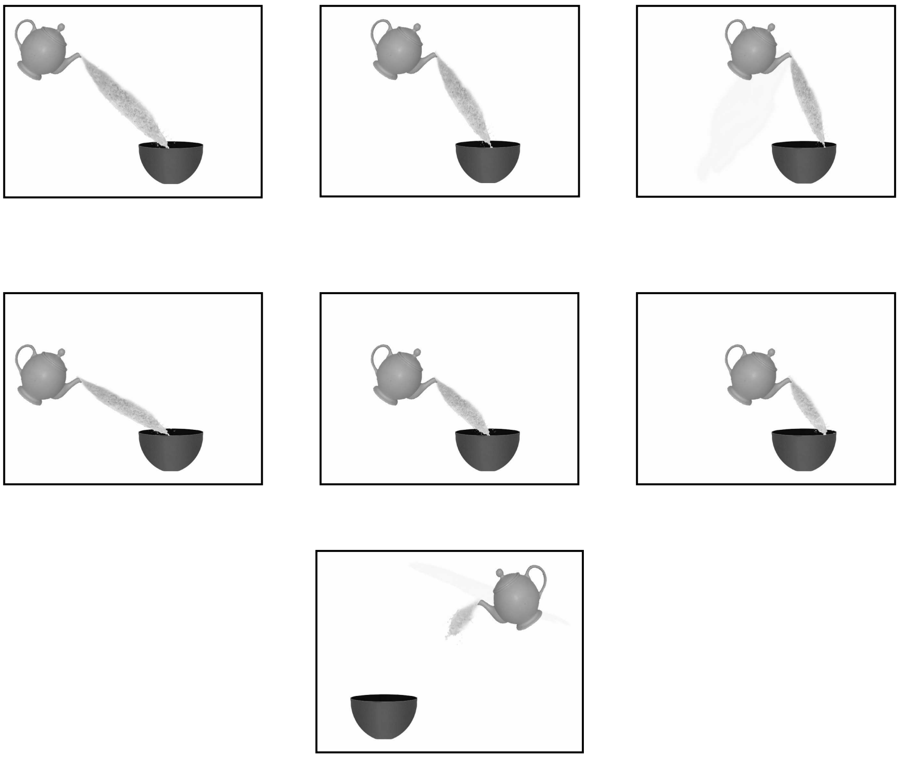
2 Implementing the Functional Geometric Framework

2.1 Experimental Data

The model we outline shortly can deal with a range of prepositions, but here we focus on *over/under/above/below*. We conducted a series of experiments (see Coventry, Cangelosi et al., in preparation, for more details) involving three different reference objects (a plate, a dish and a bowl) pre-tested in a sorting task and a rating task to be the prototypical dimensions of these objects, and a variety of other objects which were all containers (e.g., a jug). Each container was presented in each of 3 x 2 positions "higher" than the other objects (representing 3 levels of distance on the x axis and two levels on the y axis from the other object). Crucially the container was shown to pour liquid such that it ended up reaching the plate/dish/bowl (the functional condition), or missed the plate/dish/bowl (non-functional condition), or liquid was not present. Figure 2 shows some example scenes. The methodology used for these experiments involved the presentation of pictures together with sentences of the form *The located object is preposition the reference object*, and the task for participants was to rate the appropriateness of each sentence to describe each picture using a Lickert scale (range from 1 = totally unacceptable to 9 = totally acceptable).

In Experiment 1 participants saw movies of the pouring scenes (or static scenes for the no liquid condition given that no movement was involved). The results showed effects of geometry and function together with interactions between these variables and *over/under* versus *above/below*, effectively replicating the results of Coventry et al. (2001). Experiment 2 compared the full movies with just the (single frame) end states, and this established that seeing the full movie makes no difference to acceptability ratings, it is what happens to the liquid that counts. Experiment 3 then compared end states to an earlier frame in the movie showing the liquid starting to protrude from the pouring container (see bottom picture in Figure 2) in order to assess whether participants predict what will happen to the liquid in order to make judgments about the appropriateness of *over/under/above/below*. Although acceptability ratings were overall lower for the predicted scenes rather than the end state scenes, effects of geometry, function and interactions between these variables and *over/under* versus *above/below* were still present, indicating that participants do predict where the liquid will go in order to ascertain the appropriateness of these prepositions. Experiment 4 confirmed this by finding a correlation between judgments of how much of the liquid will make contact with the appropriate part of the plate/dish/bowl and acceptability ratings for *over/under/above/below*.

Fig. 2 Sample scenes used in the experiments. The top six pictures represent the 6 levels of geometry used. All six pictures show the functional condition where the liquid was shown to end up in the container. Non-functional scenes involved the same relative positions of teapot and container, but this time the liquid was shown to miss the container. The bottom picture shows an example of a scene where only the start state was shown.



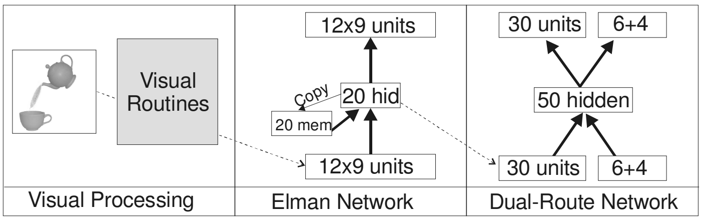
The data from these experiments indicate that participants use both information about the geometry in the scene and information about the interaction between pouring container and recipient container in the scene to assess the appropriateness of *over/under/above/below*. As has been found previously (Coventry et al., 2001), the influence of geometry was stronger for *above/below* than for *over/under*, while the influence of function (whether the liquid was shown to enter or miss the recipient container, or was predicted to enter or miss the container) was stronger for *over/under* than for *above/below*.

Data from these experiments was used as a means of testing and training the model, which we outline next.

2.2 The Computational Model

The computational model for the processing of visual scenes and the identification of the appropriate spatial preposition consists of three main modules: (1) Vision Processing, (2) Elman Network, (3) Dual-Route Network (cf. Figure 3). The first module uses a series of Ullman-type visual routines to identify the constituent objects of a visual scene (reference object, located object and liquid). The Elman network module utilises the output information from the vision module to produce a compressed neural representation of the dynamics of the scene (e.g. movement of liquid flow between the reference and located objects). This compressed representation is given in input to the dual-route (vision and language) feedforward neural network to produce a judgment regarding the appropriate spatial terms describing the visual scene. We describe each of these modules and their development in turn.

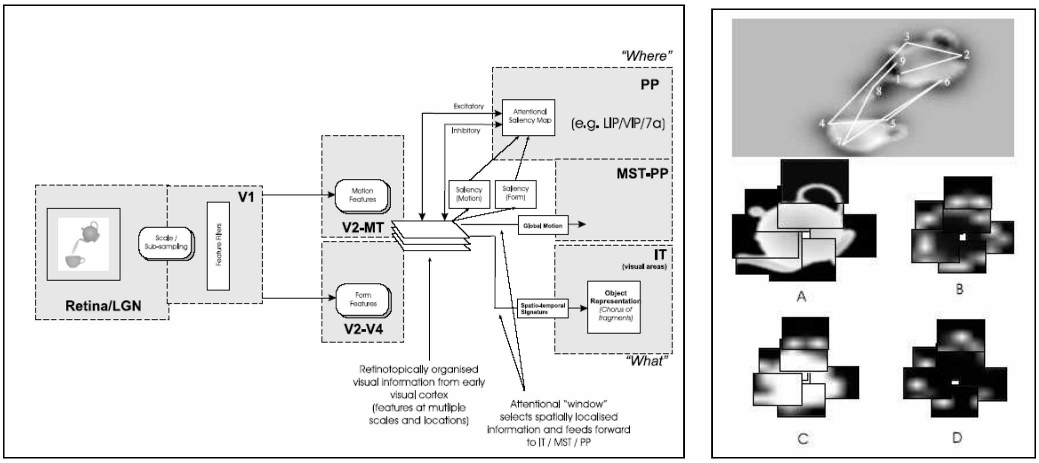
Fig. 3. Architecture of the computational model. The dotted arrows indicate functional connections between the three modules. The dual-route network has 30 visual input/output units because they copy the hidden activation of 3 different Elman networks (one with 20 hidden units, and two with 5 units each).



2.2.1 Vision processing module

In our computational model for spatial language, visual object recognition, spatial location and motion information are functionally necessary for the cognitive task. Beginning with the distinction between “what” versus “where” pathways (classically assumed to be the functionally segregated dorsal and ventral streams after Ungerleider and Mishkin, 1982), we also needed to consider the integration of object, location and motion integration when deriving a neurocomputational model. Our novel neurocomputational approach to object recognition for spatial cognition represents a compromise between the dynamic operation of the recurrent neurodynamical models of Deco and Lee (2002) for selective attention, and Edelman's (1999) feedforward chorus model for object recognition, and is conceptually congruent with Ballard et al's (1997) model (i.e. the output of our system is a plausible deictic pointer to objects in the visual scene). Image sequences (real object images composed into moving videos) are presented to the model, which processes them at a variety of spatial scales and resolutions for object form and motion features yielding a visual buffer (functionally analogous to processing in the striate visual cortex). In addition to the basic scale representation, texture, edge and region boundary features are extracted. Motion cells (in the magnocellular pathway) are modeled as uni-directional brightness gradient-sensitive cells whose outputs are combined. This is outlined in Figure 4.

Fig. 4. Left: Constituents of the Vision Processing Module and their relationships with known neural substrates. Right (Top): Snapshots of the overall saliency map after 9 fixations. Right (Bottom): Multiple Fragments of Teapot Object (A) Full visual buffer (B) Edges (C) Region/Boundary and (D) Texture


"Where"
PP
Excitatory
Inhibitory
Attentional Saliency Map
(e.g. LIP/VIP/7a)
MST-PP
Saliency (Motion)
Saliency (Form)
Global Motion
V1
Motion Features
V2-MT
Scale / Sub-sampling
Feature Filters
Retina/LGN
IT
(visual areas)
Form Features
V2-V4
Spatio-temporal Signature
Object Representation (Chorus of fragments)
"What"
Retinotopically organised visual information from early visual cortex (features at mutliple scales and locations)
Attentional "window" selects spatially localised information and feeds forward to IT / MST / PP
1
2
3
4
5
6
7
8
9
A
B
C
D

The attentional saliency map (Figure 4, Right) is a very low resolution (retinotopic) array of neurons which receive bottom-up activation from the static and motion features in the visual buffer, but which can be strongly inhibited when the region they code for is attended to or when object recognition is strong enough to require little further processing of a region. This represents information integration that might take place involving the kinds of information processed in the posterior parietal cortex. This is used to direct attention and once a region is selected (analogous to a kind of spotlight of attention), the higher-resolution information contained in the visual buffer is allowed to feedforward to the object recognition stream. Since attention selects only a windowed region of the whole visual buffer for processing in IT, our system represents a chorus of object fragments. We use Gaussian adaptive resonance models to learn the space of fragments for each object (Williamson, 1996), leading to a probabilistic implementation.

We elaborate on the visual processing and selective attention mechanism and its role in a novel chorus of fragments framework for object recognition elsewhere (Joyce et al. 2002). We show how this may form part of a larger system for spatial language comprehension and speculatively for prefrontal cortex short term visual memory and object-place binding (via the perirhinal – entorhinal – hippocampal network), all of which further ground the understanding of the visuo-spatial processing in a computational framework.

2.2.2 Elman network module

This module consists of a predictive, time-delay connectionist network similar to Elman's (1990) simple recurrent network, which we refer to hereafter as the Connectionist Perceptual Symbol System Network (CPSSN; Joyce et al., 2003). Figure 3, middle image, shows the CPSSN network as an Elman SRN. As a suitable (and plausible) input representation for the CPSSN, we propose a "what+where" code (see also Edelman, 2002). That is, the input consists of an array of some 9x12 activations (representing retinotopically organised and isotropic receptive fields) where each activation records some visual stimulus in that area of the visual field. This is the output information produced by the Vision module. In addition to the "field" representation, we augment a distributed object identity code. These codes were produced by an object representation system (Joyce et al. 2002; based on Edelman's (1999) theory) using the same videos. The CPSSN is given one set of activations as input which feedforward to the hidden units. In addition, the previous state of the hidden units is fed to the hidden units simultaneously (to provide a temporal context viz. Elman's (1990) SRN model). The hidden units feedforward producing an output which is a prediction of the next sequence item. Then, using the actual next sequence item, back propagation is used to modify weights (see Figure 3) to account for the error. The actual next sequence item is then used as the new input to predict the subsequent item and so on. Using the coding scheme discussed, we have a total input vector of length 116 (where 8 of these 116 elements code for each object, e.g. liquid, bowl, cup etc.). The output is similarly dimensioned, and there were 20 hidden units (and 20 corresponding time-delayed hidden state nodes) to represent movement of the liquid.

The network training regime was as follows: a collection of sequences are shown to the network in random order (but of course, the inputs within a sequence are presented one after another). Each sequence contains a field and object code for the "liquid" in the videos. Multiple CPSSN networks would be required to account for the other objects in the scenes. A root-mean-square error measure is used to monitor the network's performance, and the ordering of sequences is changed each time (to prevent destructive interference between the storage of each sequence). Initially, the network is trained with a learning rate of 0.25, and after the RMS error stabilises, this is reduced to 0.05 to allow finer modifications to weights. For 6 sequences, a total of about 150 presentations are required (each sequence is therefore presented 25 times) to reduce RMS averaged over the whole training set from around 35 to around 0.4.

It is quite obvious that this network is hetero-associating successive steps in the sequence of fields, but in addition, the network is performing compression and redundancy reduction (in the hidden layer) as well as utilising the state information in the time-delayed state nodes. It is also coding for the changes between sequence items (e.g. the dynamics of how the object moves over time) rather than coding individual sequence items (which would be auto-association). The model embodies the idea that representation is inherently dynamic (cf. Freyd, Pantzer & Cheng, 1988). The network should, naturally, be able to make a prediction about a sequence given any item in the sequence. Intuitively, the network should be capable of this in the case where a cue is the first item of a sequence, since the time-delayed state is irrelevant (i.e., there can be no temporal context accumulated in the time-delay nodes). However, we propose that the network is a mechanism for implementing perceptual symbols, and therefore, a requirement is that it can "replay" the properties of the visual episode that was learned. Given a cue, the network should produce a prediction, which can be fed-back as the next input to produce a sequence of "auto-generated" predictions about a sequence (viz, a perceptual symbol). Indeed, this network is able to predict the final outcome of the visual scenes (Joyce et al., 2003).

2.2.3 Dual-route network

The dual-route network is a feedforward neural network (3-layer perceptron) that receives in input the grounded "visual" information (hidden activations of the Elman networks) and linguistic data (name of located object, name of reference object, name of liquid + 4 spatial prepositions *over, above, below, under*). In output it must reproduce (auto-associate) the same visual data, and produce the names of object, which are directly grounded in the input visual data. In addition, the four output units for the spatial prepositions will encode the rating values given by subjects. This architecture is directly inspired by dual-route networks for the grounding of language (Plunkett et al., 1992; Cangelosi et al., 2000; Cangelosi in press).

This network is trained via the error backpropagation algorithm. The training and test sets consist of the 216 scenes. These are the same as those used in the experiment on the rating of *over, above, under, below* (Experiment 11 above). Of these stimuli, 195 are used for the training and 21 for the generalisation test. The overall objective of the training is that the network must learn to produce the same average ratings for the four prepositions. We did not use the average ratings as the teaching input, because this was against the principle of mutual exclusivity (Markmann 1987). During standard backpropagation training, the use of the ratings as teaching input assumes that the same scene must be simultaneously associated to the use of all four prepositions (each with an activation value proportional to the subjects' average rating). Instead, during developmental learning subjects tend to choose only one preposition to describe a scene. Naturally, the probability of choosing one preposition to describe a spatial relation is correlated to its level of appropriateness (i.e. similar to ratings). Therefore, to simulate such a learning strategy better, the original ratings of each scene-preposition pair were converted into frequency of presentation of a stimulus with an associated localist teaching input (where the output unit of the chosen preposition is 1 and the other three units are 0). To obtain such a frequency, the original average ratings were scaled and normalised within each scene and also within the whole training set. For example, individual prepositions' ratings of 7.08 (above), 7.12 (below), 3.96 (over), 4.32 (under) respectively correspond to presentation frequencies of 28, 28, 7 and 9. The conversion of ratings into preposition resulted in an epoch of 2100 stimuli.

Three networks were trained using different initial random weights and different random sets of 21 generalisation test stimuli[1]. The training parameters included a learning rate of 0.01 and momentum of 0.8, and a total number of training epochs of 500. The average final error (RMS) for the 30 vision units was 0.008 for both training and testing data, and 0.003 for the 6 output units of the object names. More importantly, for the 4 spatial preposition output units, the error was 0.044 with training data and was 0.05 with generalisation data. The error values in the preposition units were calculated off-line by comparing the actual output of the 4 preposition units and the rating data (from Experiment 3 overviewed above) converted to produce the stimulus frequencies (the actual error values used for the weight correction are always higher because they use localist teaching input). These results clearly indicate that the networks produce rating values similar to that of experimental subjects. They also indicate that the training algorithm based on presentation frequency, instead of rating teaching input, works well and provides a psychologically-plausible learning regime.

2.4 Interplay between Experimental and Computational Work

The development of the computational model has been conducted in parallel with experimental investigations. However, in the early part of the development of the model, the experimental work has mostly influenced the model design. For example, in the previous section we explained that the training/test stimuli and the rating values were directly taken from one experiment. Later on in the development of the model, it was the model that directed some of the directions and objectives of the experimental investigation. In particular, new simulations produced some predictions that were subsequently tested in new experiments.

Research on the design and test of the Elman module had shown that these networks were able to predict and auto-generate the final outcome of the visual scenes, once they were given an initial cue (e.g few initial frames). The network would produce the next prediction frames, which were fed-back as the next input. To integrate such prediction ability in the overall spatial language model, the hidden activation values of these auto-generated sequences were used as visual input of the dual-route network. The model was then run as usual to produce the ratings of the 4 prepositions.

To establish if the new ratings provided by the model were consistent with those produced by real subjects, a new experiment was conducted (Experiment 4, see

0

[1] Here we report only the data from the best simulation. Different parameters values and hidden layer sizes were tested.

above). The results for this experiment, together with the results of Experiment 3, strongly suggest that subjects had to mentally "play" the visual scene and autogenerate the outcome of the scene to rate the linguistic utterance. This is very similar to what the model does, when the Elman network autogenerates the visual scene, and the dual-route network uses the Elman net's activations to produce new ratings. The Elman network used the first 3 out of 7 frames. This corresponds to the frames 0, 10 and 20 (Elman networks only see a frame every 10). The comparison of the subjects' rating data and the networks' output of the 4 prepositions resulted in an RMS error of 0.051. This is a very low error level, and confirms that the model had predicted very accurately the ratings. Overall, this result and those on the dual-route tests support the development of a psychologically-plausible model for spatial language.

3 Discussion: Extension and Links

The model we have outlined has been tested across other spatial relations as well as *over/under/abive/below*, including the importance of location control for the preposition *in*. Currently we are extending the model so that it can return a description of the number of objects in the visual input scene as well as the spatial relations between objects depicted. Vague quantifiers like *a few* and several exhibit many of the same context effects that have been observed for spatial prepositions. For example, relative size of figure and ground objects (Hormann, 1983; Newstead & Coventry, 2000) and expected frequency (Moxey & Sanford, 1993) have both been shown to affect the comprehension of quantifiers; *A few cars* is associated with a smaller number than *a few crumbs*, and *some people in front of the cinema* is associated with more people than *some people in front of the fire station*. These context effects appear very similar to the range of effects in evidence for spatial preposition. Therefore the issue we are exploring is that these context effects originate from visual processing constraints such that information regarding specific numbers of objects in a scene cannot be derived very easily from visual processing of that scene.

From a theoretical perspective the functional geometric framework and the implementation of it are consonant with recent develops in the embodied cognition literature. The idea that meaning is constructed as a result of putting together multiple constraints fits with recent work by Glenberg and colleagues (e.g., Glenberg, 1997; Glenberg & Kaschak, 2002) and by Barsalou (e.g., Barsalou, 1999). Glenberg and colleagues have proposed that the meaning of a sentence is constructed by indexing words or phrases to real objects or perceptual analog symbols for those objects, deriving affordances from the objects and symbols and then meshing the affordances under the guidance of syntax. Barsalou (1999) also places similar emphasis on perceptual representation for objects and nouns in his perceptual symbol systems account. For Barsalou, words are associated with schematic memories extracted from perceptual states which become integrated into what Barsalou terms *simulators* (see also Grush, in press). As simulators for words develop in memory, they become associated with simulators for the entities and events to which they refer. Furthermore, once simulators for words become linked to simulators for concepts, Barsalou argues that words can then control simulations. We hope to be able to extent the model further by also considering interaction with objects by the model more directly (e.g., through the addition of a robotic arm), rather than simply observing interactions between objects. We hope that such developments help move embodiment arguments from the theoretical arena to showing how these ideas can be realized in a working neuro-computational model.

References

Ballard, D. H., Hayhoe, M. M., Pook, P. K. & Rao, R.P.N. (1997) Deictic codes for the embodiment of cognition. *Behavioural and Brain Sciences, 20*, 723-767.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22(4),* 577-660.

Cangelosi (in press). Grounding symbols in perceptual and sensorimotor categories: Connectionist and embodied approaches. In H. Cohen & C. Lefebvre (Eds), *Categorization in Cognitive Science*, Elsevier

Cangelosi A., Greco A. & Harnad S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science, 12(2),* 143-162.

Carlson-Radvansky, L.A., Covey, E. S. & Lattanzi, K. M. (1999). “What” effects on “where”: Functional influences on spatial relations. *Psychological Science, 10(6),* 516-521.

Carlson-Radvansky, L. A., & Radvansky, G. A. (1996). The influence of functional relations on spatial term selection. *Psychological Science, 7,* 56-60.

Carlson-Radvansky, L. A., & Tang, Z. (2000). Functional influences on orienting a reference frame. *Memory & Cognition*, 28, 812-820.

Coventry, K. R. (1998). Spatial prepositions, functional relations and lexical specification. In P. Olivier and K.-P. Gapp (Eds.) *Representation and Processing of Spatial Expressions,* (pp. 247-262). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Coventry K.R., Cangelosi A. et al. (in submission). Spatial language and perceptual symbol systems: Implementing the functional geometric framework.

Coventry, K. R., Carmichael, R. & Garrod, S. C. (1994). Spatial prepositions, object-specific function and task requirements. *Journal of Semantics, 11,* 289-309.

Coventry, K. R. and Garrod, S. C. (2004). *Saying, Seeing and Acting. The Psychological Semantics of Spatial Prepositions.* Lawrence Erlbaum Associates. Essays in Cognitive Psychology Series.

Coventry, K. R., & Garrod, S. C. (in press). Spatial prepositions and the functional geometric framework. Towards a classification of extra-geometric influences. In L. A. Carlson, & E. van der Zee (Eds.), *Functional features in language and space: Insights from perception, categorization and development.* Oxford University Press.

Coventry, K. R. & Mather, G. (2002). The real story of ‘over’?. In K. R. Coventry & P. Olivier (Eds.), *Spatial Language: Cognitive and Computational Aspects*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Coventry, K. R. & Prat-Sala, M. (2001). Object-specific function, geometry and the comprehension of ‘in’ and ‘on’. *European Journal of Cognitive Psychology, 13(4),* 509-528.

Coventry, K. R., Prat-Sala, M. & Richards, L. V. (2001). The interplay between geometry and function in the comprehension of ‘over’, ‘under’, ‘above’ and ‘below’. *Journal of Memory and Language, 44,* 376-398.

Deco, G. and T.S. Lee (2002) A Unified Model of Spatial and Object Attention Based on Intercortical Biased Competition. In Press, Neural Computation.**UPDATE**

Edelman, S. (1999). *Representation and Recognition in Vision.* MIT Press, 1999.

Edelman, S. (2002). Constraining the Neural Representation of the Visual World. *Trends in Cognitive Sciences, 6,* 125-131

Elman, J.L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211

Feist, M.I., & Gentner, D. (1998). On Plates, Bowls, and Dishes: Factors in the Use of English 'in' and 'on'. Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, 345-349.

Ferenz, K. S. (2000). The role of nongeometric information in spatial language. PhD Thesis, Dartmouth College, Hanover, New Hampshire.

Freyd, J., Pantzer, T., & Cheng, J. (1988). Representing statics as forces in equilibrium. Journal of Experimental Psychology: General, 117, 395-407.

Gapp, K. P. (1995). Angle, distance, shape and their relationship to projective relations. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the* Cognitive Science Society (pp. 112-117). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Garrod, S., Ferrier, G., & Campbell, S. (1999). In and on: Investigating the functional geometry of spatial prepositions. Cognition, 72, 167-189.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences, 20(1),* 1-55.

Glenberg, A. M., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin and Review, 9(3), 558*-565.

Grush, R. (in press). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences.*

Herskovits, A. (1986). *Language and Spatial Cognition. An interdisciplinary study of the prepositions in English.* Cambridge, UK: Cambridge University Press.

Hormann, H. (1983). Then calculating listener, or how many are einige, mehrere and ein paar (some, several and a few). In R. Bauerle, C. Schwarze, & A, von Stechow (Eds.), *Meaning, use and interpretation of language*. Berlin; De Gruyter.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17(2),* 187-245.

Jeannerod, M.(2001) Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage, 14*, S103-S109.

Joyce D., Richards L., Cangelosi A., Coventry K.R. (2002), Object representation-by-fragments in the visual system: A neurocomputational model. In L. Wang, J.C. Rajapakse, K. Fukushima, S.Y. Lee, X. Yao (Eds), *Proceedings of the 9th International Conference on Neural Information Processing (ICONP02)* IEEE Press, Singapore.

Joyce. D. W., Richards, L. V., Cangelosi, A. & Coventry, K. R. (2003). On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. In F. Dretje, D. Dorner & H. Schaub (Eds.), The Logic of Cognitive Systems. *Proceedings of the Fifth International Conference on Cognitive Modelling*, pp147-152. Universitats-Verlag Bamberg, Germany.

Landau, B., & Jackendoff, R. (1993). 'What' and 'where' in spatial language and cognition. *Behavioural and Brain Sciences, 16(2),* 217-265.

Logan, G. D., & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and Space* (pp. 493-530). Cambridge, Mass.: MIT Press.

Markman E.M. (1987), How children constrain the possible meanings of words," In Concepts and conceptual development: Ecological and intellectual factors in categorization. Cambridge University Press.

Moxey, L. M., & Sanford, A. J. (1993). *Communicating Quantities. A Psychological Perspective*. Lawrence Erlbaum Associates; Hove, East Sussex.

Newstead, S. E., & Coventry, K. R. (2000). The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology*, 12(2), 243-259.

Plunkett, K., Sinha, C., Moller, M.F & Strandsry, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary grouth in children and a connectionist net. *Connection Science, 4(3-4*), 293-312.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism.* Cambridge Mass.: MIT Press.

Regier, T., & Carlson, L.A. (2001) Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General, 130(2),* 273-298.

Schwartz, D. L. (1999). Physical imagery: Kinematic versus dynamic models. *Cognitive Psychology, 38*, 433-464.

Talmy, L. (1983). How language structures space. In H. Pick, & L. Acredolo (Eds.), *Spatial Orientation: Theory, research and application*, (pp. 225-282). New York: Plenum Press.

Ullman, S. (1984). Visual routines. *Cognition, 18,* 97-159.

Ullman, S. (1996). *High-level Vision. Object recognition and visual cognition*. Cambridge, MA; MIT Press.

Ungderleider, l, & Mishkin, M. (1982). Two cortical visual systems. In D. Ingle, M. Goodale, & R. Mansfield (Eds.), *Analysis of Visual Behaviour*. Cambridge, MA: MIT Press.

Williamson J.R. (1996). “Gaussian ARTMAP: A Neural Network for Fast Incremental Learning of Noisy Multidimensional Maps”, *Neural Networks, 9(5),* 881-897.

Zee, E. van der, & Watson, M. (in press). Between function and space: How functional and spatial features determine the comprehension of “between”. In L. A. Carlson, & E. van der Zee (Eds.), *Functional features in language and space: Insights from perception, categorization and development.* Oxford University Press.