GROUNDING LANGUAGE IN PERCEPTION: A CONNECTIONIST MODEL OF SPATIAL TERMS AND VAGUE QUANTIFIERS

ANGELO CANGELOSI[A] KENNY R. COVENTRY[B] ROHANA RAJAPAKSE[A]
DAN JOYCE[C] ALISON BACON[B] LYNN RICHARDS[B] STEVEN N. NEWSTEAD[B]

[A] *Adaptive Behaviour and Cognition Group*
School of Computing, Communications and Electronics, University of Plymouth,
Drake Circus, Plymouth PL4 8AA, UK
{acangelosi, rrajapakse}@plymouth.ac.uk

[B] *School of Psychology, University of Plymouth*
{kcoventry, ambacon, lvrichards }@plymouth.ac.uk

[C] *Sobell Dept. of Motor Neuroscience,University College London,*
dan.joyce@virgin.net

This paper presents a new connectionist model of spatial language based on real psycholinguistic data. It puts together various constraints on object knowledge ("what") and on object localisation ("where") in order to influence the comprehension of a range of linguistic terms, mirroring what participants do in experiments. The computational model consists of a vision processing module for input scenes, an Elman network module for the representation of object dynamics, and a dual-route network for the production of object names and linguistic prepositions describing the scene. Preliminary simulations on the prediction of spatial term ratings are presented, and extensions of the model to vague quantifiers and other syntactic categories are considered.

1. Introduction

Spatial language and cognition is concerned with the understanding of the cognitive and linguistic mechanisms affecting the appraisal of spatial tasks. One very important issue is the relative extent to which various constraints on object knowledge ("what" information) and on object localisation ("where" information) influence the comprehension of a range of linguistic terms. Extensive research has been dedicated to the understanding of the effects of "where" information, that is geometric factors related to the relative position and orientation of the objects, in the use of spatial prepositions. For example, Logan and Sadler (1996) have proposed the existence of individual spatial templates for prepositions such as *above* and *under*. These templates specify the level of appropriateness of each spatial term for the different positions (in a 7x7 grid) of a located object with respect to a reference point. Regier (1996) developed a

constrained connectionist model of spatial language able to deal with the effects of geometric factors on various spatial terms. More recently, Regier and Carlson (2001) have studied the combined effects of attentional and geometric factors in their Attention Vector Sum (AVS) model. In parallel with these studies on geometric factors, significant evidence has been gathered on the role of a range of "extra-geometric" factors on spatial language comprehension. These extra-geometric factors include the knowledge of object properties and function, and general knowledge of object interaction dynamics. For example, Coventry et al. (2001) have investigated the relative importance of object function and geometric position on the comprehension of *over*, *under*, *above* and *below*. Whether an object is depicted as fulfilling its function or not (e.g. rain shown to fall on an umbrella protecting a man from getting wet, or shown to miss the umbrella and therefore failing to protect the man) is a better predictor of the acceptability of *over* and *under* to describe such scenes than the relative positions of umbrella and man, for example, while conversely the relative positions of umbrella and man are a better predictor the acceptability of *above* and *below* than function.

More recently, Coventry and Garrod (2004) have developed the "functional geometric framework" to explain the integration of the "what" and "where" factors in spatial language. They argue that the application of geometric and extra-geometric routines underlie the comprehension of spatial prepositions. The application of such 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. Such a framework is also consistent with growing theoretical arguments and experimental evidence on the role of grounding language in action and perception (Barsalou 1999; Glenberg & Kashak 2002; Glenberg, this volume; Cangelosi, in press). For example, the idea that meaning is constructed as a result of putting together multiple constraints fits with recent work by Glenberg and Kaschak. They 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) 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. Cangelosi (in press) uses the Cognitive Symbol Grounding framework based on the hypothesis that symbols are directly grounded in internal categorical representations, whilst

at the same time having logical (e.g. syntactic) relationships with other symbols. Some symbols, those corresponding to the basic vocabulary, need to be learned and directly grounded, through experience, in the objects they refer to (and the categorical representations that they activate). This is the case, for example, of words learned during early lexical development. Other symbols can then be grounded in representations of categorical entities constructed by the individuals, not necessarily through direct experience and interaction (e.g. when new concepts are learned through deduction). The internal categorical representations, constituting the meanings upon which symbols are grounded, include perceptual, sensorimotor, and social categories, as well as internal state representations. He considers two modelling approaches to symbol grounding: (i) the connectionist approach, based on artificial neural networks for category learning and naming tasks, and (ii) the embodied modelling approach, based on adaptive agent simulations and cognitive robots. These models provide an integrative view of cognitive systems and help our understanding of the relationships between vision, action and language.

In this paper we present a new computational model for spatial language, in which visual scenes are described by selecting the spatial terms that most appropriately describe them (i.e. consistent with human subjects' acceptability ratings). Due to the fact that the model is able to ground the selection of the spatial terms directly into the visual scene, this work helps bridge the gap between theories of meaning which capture meaning in terms of symbol-symbol relations (Landauer & Dumais, 1997) versus those which "ground" language directly in perceptual representation (Regier and Carlson, 2001). In particular, the new model will be able to directly ground the names of objects in visual scenes. Spatial terms, such as the prepositions *over*, *under*, *above* and *below* will also be grounded in information on objects' locations and interaction provided in the input scenes. This will also create a system in which symbol-symbol relationships (e.g. between prepositions and nouns) also permit a prediction of the interaction between objects.

2. The Computational Model

The computational model has a hybrid vision-connectionist architecture (Figure 1) with three main modules: (1) Visual Routines, (2) Elman Networks, (3) Dual-Route Network.

The Visual Routine module uses a series of Ullman-type vision processing routines (Joyce et al. 2002) to identify the constituent objects of a visual scene. It is directly inspired by recent findings and theories of visual object processing,

such as Edelman's (1999) feedforward chorus model. The input to the Visual routine module consists of seven frames from 60-second movies (one frame every ten seconds). The scene involves three objects: a located object (e.g. teapot), a liquid substance (e.g. water) and a reference object (e.g. a container such as a cup). The frames 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. In addition to the basic scale representation, texture, edge and region boundary features are extracted. The processing of each frame results in three arrays of 9x12 activations, representing retinotopically organised and isotropic receptive fields for each of the three objects.

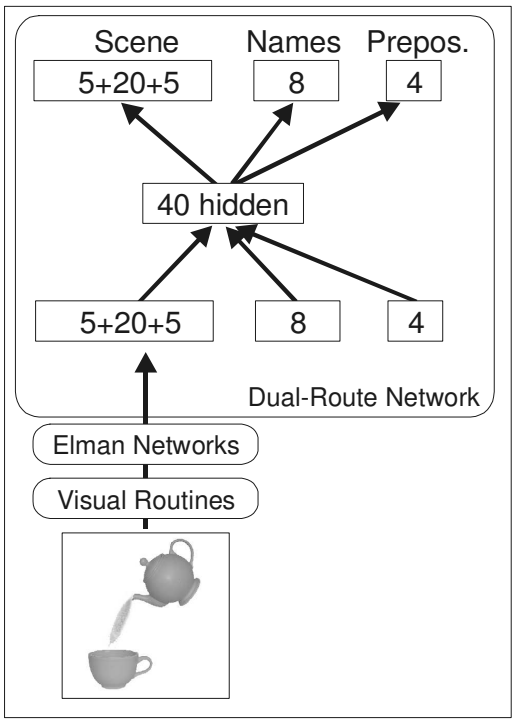

Figure 1 – Dual-route network with input from Visual Routines and the Elman Networks modules.

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 located and reference objects). Three separate Elman networks are used, respectively for each of the three object types in the scene (located object, liquid, reference objects). Three

different networks are currently needed, because the vision module is not currently able to segment and identify objects autonomously. All networks have 108 input and output units. The networks perform a typical prediction task, by producing in output the matrix of the next frame in the 7-frame sequence. The Elman network predicting the liquid flow has 20 hidden units and 20 context (memory) units. The two networks for the located and reference objects only require 5 hidden and context units, because the prediction task is trivial (objects are static). The scope of the Elman network module is to compress the temporal and dynamic information in the scene. This is achieved in the hidden units' activation at the last frame presentation, when the networks have been able to predict all the frames of the scene. The network training protocol consists of a collection of sequences shown to the network in random order (but with fixed sequential order for the seven frames of each scene). Networks are able to correctly learn the prediction task for both training and generalization scenes (Joyce et al., 2003).

The third module consists of a Dual-Route neural network. This architecture combines visual and linguistic information for both linguistic production and comprehension tasks (Plunkett et al., 1992; Cangelosi et a., 2000). This is the core component of the model, as it integrates visual and linguistic knowledge to produce a description of the visual scene. The network receives in input information on the scene through the activation values of the Elman networks' hidden units. It will then produce in output a judgment regarding the appropriate spatial terms describing the visual scene and the names of the objects involved in the scene. The activation values of the linguistic output nodes correspond to rating values given by subjects for the spatial prepositions.

The network used in this simulation had 30 input visual units and 12 input linguistic nodes. The 30 visual nodes corresponded to the 30 hidden units of the three Elman networks (respectively 5 for the network processing the teapot, 20 the liquid and 5 the containers).The linguistic units consisted of the 8 names of objects and the 4 prepositions *over*, *under*, *above* and *below*. The dual route network has 40 hidden nodes. The output layer had the same number and type of units as those in input. As a matter of fact, the dual route network can be considered to be an autoassociator.

3. Simulation and Results

3.1.*Experimental Data and Training of the Network*

Preliminary simulations of the model focussed on the spatial prepositions *over*, *under*, *above* and *below*. For the model training, experimental data from a series of experiments on spatial language were used (Coventry et al., submitted). The visual scenes, used for both the experiment and the model, involved: 3 different reference objects (containers: a plate, a dish and a bowl), 2 levels of closure of container (lid on/closed and lid off/open), 6 different positions for the located objects (in a 3x2 grid position "higher" than the other objects), 2 directions of the located object (left and rightward facing), and 3 functional conditions (the liquid ends in the contained, or misses it, or no liquid is present). This constitutes a stimulus set of 216 movies in total (i.e. 3x2x6x2x3 experimental design). Figure 2 shows the initial and final frames of a sample scene.

The methodology used for these experiments involved the presentation of pictures together with sentences of the form "*The teapot is over the cup*". Participants had to rate the appropriateness of each sentence to describe the movies using a Lickert scale (range from 1 = totally unacceptable to 9 = totally acceptable). Typically, these experiments show effects of geometry and function, together with interactions between these variables and *over/under* versus *above/below* (see also Coventry et al., 2001).

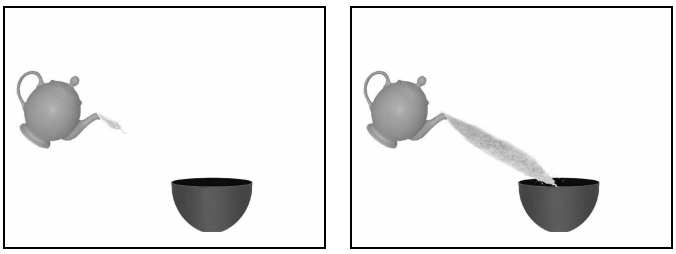

Figure 2. Example of training stimuli. First (left) and last (right) frame of the 60 second movie. This represents a functional scene, i.e. when the liquid finishes in the container.

The subjects' ratings were used to train the dual-route neural network through error backpropagation. An innovative method for backpropagation training was used, where ratings were converted into probability of scene-preposition pairings. This is to avoid a training regime in which the network is assumed to be giving four simultaneous judgments for each scene. Instead

human subjects, especially during developmental learning, tend to choose only one preposition to describe a scene (although the choice of a preposition may depend on various contextual factors). To simulate such a learning strategy, the original ratings of each scene-preposition pair were converted into frequency of presentation of a stimulus with an associated localist teaching input. At each learning cycle, only one output preposition unit is set to 1 (maximum activation/choice), whilst the other 3 prepositions units are trained to be inactive. To obtain such a frequency, the original average ratings were scaled and normalized within each scene and within the whole training set. For example, two quite extreme ratings of 7.12 and 3.96 (range 1-10) respectively correspond to presentation frequencies of 28 and 7.

The conversion of ratings into preposition frequencies resulted in a training epoch of 21611 stimuli (scene-preposition pairs) for the dual-route network. Three networks were trained using different initial random weights and different random sets of generalisation test stimuli (10% of scenes that were never used during training). The training parameters included a learning rate of 0.01 and momentum of 0.8, and a total number of training epochs of 500.

3.2.*Results*

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 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 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. The networks' ability to correctly generalize the use (ratings) of the four prepositions with the novel scenes of the generalization set indicates that the model has learned the relationships between the objects involved in their scene, their geometrical and functional properties and the linguistic terms. 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.

Additional simulations (Coventry et al., in submission) have also shown that the model can accurately predict new experimental rating data for new scenes. For example, this is the case where only the initial frames are shown and the

networks (i.e. the Elman nets) must "replay" the scene and predict its end frame (i.e. where the liquid ends). The Elman network hidden activation values were then passed to the dual route network to generate new rating for the four spatial prepositions. To compare the model's prediction with the performance of real subjects, a new experiment was conducted. Subjects had to predict the end states of the initial frames of movie and rate the appropriateness of the linguistic descriptions. The acceptability ratings for both networks and subjects were overall lower for the predicted scenes rather than the end state scenes. The dual rout network error for the four preposition units was below 5%. In addition, in both networks and subjects the 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.

4. Discussions and Future Work

The preliminary results of this hybrid vision and connectionist model of spatial prepositions demonstrated the feasibility of building a language processing model directly grounded in perception. In addition, the model's ability to replicate accurately subjects' performance in the liquid-flow prediction simulation (and experiment) further supports its psychological validity.

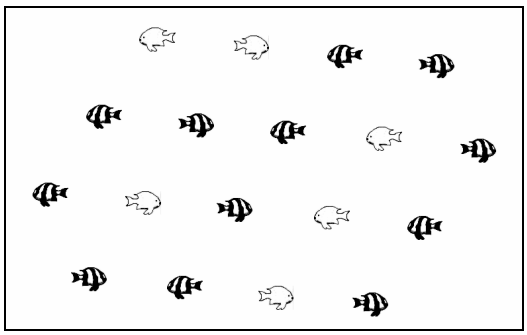

Figure 3. Example of stimuli for quantifier experiments. Subjects (and networks) have to rate sentences such as "There are *several* striped fish.

This model is currently being extended to deal with further linguistic terms, namely vague quantifiers such as *some*, *few*, *a few*, *lots of* (Figure 3). The hypothesis is that this grounded connectionist approach will permit the identification of the main mechanisms responsible for quantification judgment and their linguistic expression. 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 objects and their expected frequency (e.g. Hormann 1983; Newstead & Coventry 2000) have both been shown to affect the comprehension of quantifiers. "A few cars" is associated with a smaller number than "a few crumbs". In addition, new experiments (Coventry et al., in preparation) have demonstrated that the rating of vague quantifiers is affected by the extent to which objects are grouped together and the degree of spacing between objects. The issue we are exploring with the new model 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.

We also hope to be able to extend the model further by considering the direct interaction of the model (e.g. a cognitive agent) with objects in its environment. This is in contrast with the use of a "passive" model that observes interactions between objects through the presentations of movies. This new model might involve the use of a robotic arm (Massera et al., this volume; Cangelosi in press) that builds categorical and linguistic representation of the world by learning to manipulate objects and interact with them. This approach is consistent with the embodied framework in cognitive psychology (e.g. Barsalou's and Glenberg's grounded theories of cognition) and in cognitive systems studies (e.g. Steels, 2003).

Acknowledgments

This research was supported by the UK Engineering and Physical Research Sciences Council (EPSRC Grants GR/N38145 and GR/S26569)

References

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.

Coventry, K. R. & Garrod, S. C. (2004). *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Psychology Press, Hove.

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.

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

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

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

Joyce D., Richards L., Cangelosi A., Coventry K.R. (2002), Object representation-by-fragments in the visual system: A neurocomputational model. In L. Wang et al. (Eds), *Proceedings of the 9th International Conference on Neural Information Processing (ICONP02)* IEEE Press.

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 et al. (Eds.), *The Logic of Cognitive Systems. Proceedings of the Fifth International Conference on Cognitive Modelling*, pp147-152. Universitats-Verlag Bamberg, Germany.

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.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.

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.

Massera G., Nolfi S., Cangelosi A. (in press), Evolving a simulated robotic arm able to grasp objects. In A. Cangelosi A., G. Bugmann & R. Borisyuk (Eds.), *Modelling Language, Cognition and Action: Proceedings of the 9th Neural Computation and Psychology Workshop*. Singapore: World Scientific

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 growth 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.

Steels L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Science*. 7(7), 308-312.