Autonomous Neural Dynamics to Test Hypotheses in a Model of Spatial Language

Mathis Richter (mathis.richter@ini.rub.de)
Jonas Lins (jonas.lins@ini.rub.de)
Sebastian Schneegans (sebastian.schneegans@ini.rub.de)
Yulia Sandamirskaya (yulia.sandamirskaya@ini.rub.de)
Gregor Schöner (gregor.schoener@ini.rub.de)

Institut für Neuroinformatik, Ruhr-Universität Bochum, 44870 Bochum, Germany

Abstract

Resolving relational spatial phrases requires that a coherent mapping emerges between a visual scene and a triad of two objects and a relational term. We present a theoretical account that solves this problem based on neural principles. A neural dynamic architecture represents perceptual information in activation fields that make detection and selection decisions through neural interaction. Activation nodes and their connectivity to the perceptual fields represent concepts. Dynamic instabilities enable the autonomous sequential organization of the processing steps needed to resolve relational spatial phrases. These include bringing visual objects into the attentional foreground, performing spatial transformations, and making matching decisions. We demonstrate how the neural architecture may autonomously test different hypotheses to resolve relational spatial phrases. We discuss how this neural process account relates to existing theoretical perspectives and how to move beyond the entry point sketched here.

Keywords: spatial language; sequence generation; autonomy; hypothesis testing; neural dynamics; Dynamic Field Theory

Introduction

Language enables humans to communicate about shared environments. For instance, I may use language to direct your attention to an object in a visual scene. When several similar objects are visible such as in Fig. 1a, using object identity ("cup") or feature ("red") alone is not sufficient. A relational spatial phrase, for example "the red cup to the left of the green cup", resolves ambiguity in such situations. Even in the scene in Fig. 1b, in which no object can be singled out by feature reference, this phrase uniquely specifies one of them. A typical relational phrase like the one above consists of a

(a) (b)

Fig. 1: Visual scenarios affording the use of spatial language.

target (the red cup) and a reference (the green cup), relative to which a relational term (to the left) is applied. Interpreting such a phrase may require that different pairs of objects be examined. Psychophysical evidence from visual search tasks suggests that this happens in sequence rather than in parallel (Logan, 1994). Selecting the reference and target object of such a pair also appears to happen sequentially. This is suggested by characteristic shifts of attention found using EEG measurements (Franconeri, Scimeca, Roth, Helseth, & Kahn, 2012), eye-tracking (Burigo & Knoeferle, 2011), and behavioral cuing (Roth & Franconeri, 2012).

The processing steps involved in interpreting a relational spatial phrase include binding each object to its role, centering the reference frame on the reference object, mapping the spatial term onto this reference frame, and assessing the match of the target object with the spatial term (Logan & Sadler, 1996). While such discrete processing steps appear natural in information processing terms, they require an explanation in neural systems. At the population level that is relevant to behavior, neural activity evolves continuously in time. The flow of activation is determined by the structure of neural networks. Flexibility is thus an achievement in neural processing, not a given. In previous work we have provided the basis for realizing some of these processing steps in accordance with neural principles (Lipinski, Schneegans, Sandamirskaya, Spencer, & Schöner, 2012). This work is based on the framework of Dynamic Field Theory (DFT; Schneegans & Schöner, 2008), in which activation peaks are units of representation. The model addresses the attentive selection of target and reference objects and proposes a neural architecture that transforms the location of the target object into a frame centered on the reference object. Spatial terms are encoded relative to that frame as patterned neural connections. While the neural processes of bringing objects into the attentional foreground and activating spatial terms unfold autonomously, the sequential order of these different operations is controlled through signals from outside the system.

In this paper we provide a fully autonomous neural dynamic architecture that generates sequences of processing steps to interpret and generate relational spatial language. Within the framework of DFT, we take inspiration from earlier work on the autonomous generation of behavioral sequences (Sandamirskaya & Schöner, 2010; Richter, Sandamirskaya, & Schöner, 2012). The key idea is that elementary processing steps are characterized by certain aspects that can be implemented in a neural system: The neural representation of an *intention* drives activation in those neural structures that are relevant for executing the processing step. The resulting changes in activation states are detected through a *condition of satisfaction*, which indicates the successful com-

pletion of a step, or a *condition of dissatisfaction* that indicates its failure. These detection events are bifurcations of the neural dynamics and they trigger the transition to the next processing step. Detecting completion and triggering appropriate subsequent steps enables flexible control of the sequential chain of processes, so that spatial relational phrases of different structure can be resolved. Moreover, in a situation like Fig. 1b, where there are multiple eligible candidates for the roles of reference and target, being able to detect failure enables the architecture to test different hypotheses.

Methods

DFT describes neural activity at the population level through dynamic fields (DFs), activation patterns defined over continuous feature dimensions (e.g., color hue value or spatial position). DFs evolve continuously in time under the influence of external inputs and lateral interactions within the DF as described by an integro-differential equation

$$\tau \dot{u}(x,t) = -u(x,t) + h + S(x,t) + \int f(u(x',t))\, w(x-x')\, dx'.$$

Here, $u(x,t)$ is the activation field over feature dimension x at time t, τ is a time constant, h is the negative resting level, and $S(x,t)$ is external input. An output signal $f(u(x,t))$ is determined from the activation via a sigmoid function with threshold at zero. This output is then convolved with an interaction kernel w that consists of local excitation and surround inhibition (Amari, 1977).

The interaction patterns promote the formation of localized activation peaks as attractor states of the DF, which form when localized input drives activation beyond the output threshold. The peak formation constitutes an instability in the field dynamics (the detection instability), in which the subthreshold attractor state becomes unstable. Such instabilities form discrete events that emerge from the time-continuous changes of activation, and are critical in the autonomous sequential organization of neural processes.

DFs can support multiple peaks, which may be self-sustained in the absence of localized input due to self-excitation and constitute a form of working memory. Alternatively, DFs with sufficiently strong inhibitory interactions accommodate only a single peak at a time, leading to autonomous selection decisions among localized inputs. Discrete activation nodes with a neural dynamics analogous to DFs are approximate descriptions of the field dynamics around a peak location that may be 'on' (peak present) or 'off' (sub-threshold). Different DFs can be connected to form larger architectures in which the output of one field serves as input for another field, and fields of different dimensionalities may be coupled to each other along shared feature dimensions.

Architecture

The DFT architecture shown in Fig. 2 constitutes a single, high-dimensional dynamical system. Neural representations of perceptual feature spaces (right part of the figure) are combined with neural representations of concepts (left part of the figure). The concepts are implemented as synaptic connection patterns between discrete activation nodes and the perceptual feature spaces of DFs. In terms of neural grounding, these nodes are akin to amodal records of sensorimotor activation patterns in cortex (e.g., Damasio, 1989). The perceptual representations receive visual input from a camera image, and provide the substrate for instantiating the concepts and for binding them to objects in the visual scene. A subset of nodes (top left) implements the aspects of intention and condition of satisfaction for elementary processing steps. These nodes control the progression of the dynamical system through the steps by activating concepts and modulating DF activation levels. We now step through the architecture, starting with the top right of Fig. 2 and proceeding clockwise.

Perceptual system and feature attention

The visual input to the system consists of a distribution of salient colors extracted from the camera image. It is fed into the three-dimensional *perceptual field*, which forms an activation distribution over the two spatial dimensions of the image and one color dimension (top right in Fig. 2). To generate activation peaks and thereby bring specific objects into the attentional foreground, the perceptual field requires additional input from the *color intention field* (top middle in Fig. 2). This field reflects the color of a task-relevant item and projects into the perceptual field to implement a form of feature attention.

The color intention field is coupled to two more fields relevant for the sequential organization of operations. The *color condition-of-satisfaction (CoS) field* receives excitatory input from both the color intention field and the perceptual field. It forms a peak if these two inputs coincide in color space, and thereby signals that an item of the desired color has been selected in the perceptual field. Conversely, the *color condition-of-dissatisfaction (CoD) field* is inhibited by the color intention field, but has a higher resting level. Excitatory input from the perceptual field induces a peak here when an item of any non-matching color has been selected.

Representing spatial relations

The perceptual field provides purely spatial input to two fields that represent object locations for the different roles in a relational phrase. The location of a single reference object is captured in the *reference field*, and the locations of one or more potential target items are represented in the *target candidates field*. Based on the peak positions in these two fields, the relative positions of the target candidates with respect to the reference object can be determined by a reference frame transformation (blue diamond in Fig. 2). This is implemented here as a convolution of the field outputs, but may be realized neurally using a four-dimensional field (Lipinski et al., 2012). The result is fed into the two *relational fields*. There is again a CoS and a CoD field here, with roles analogous to those in the color representation. These fields receive additional input

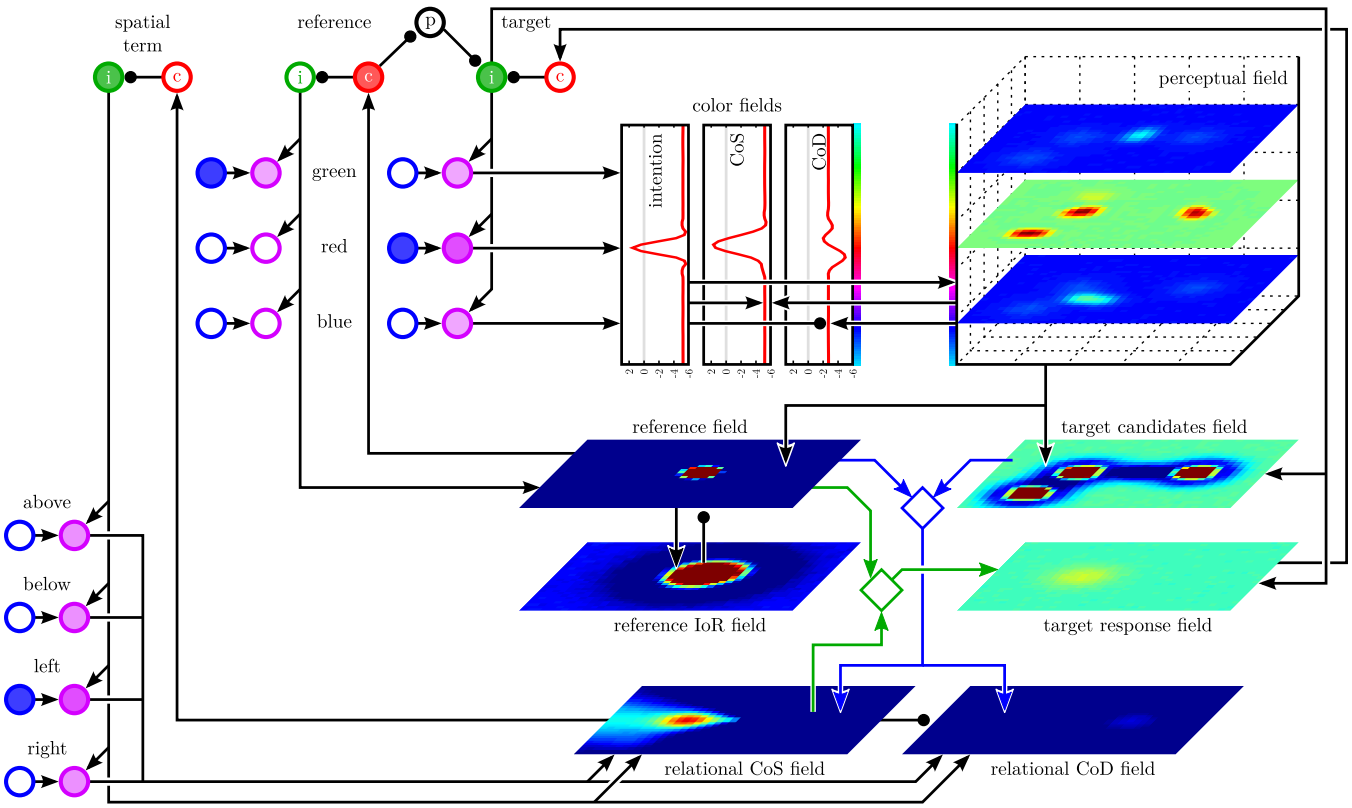

Fig. 2: Overview of the architecture for autonomous spatial language. This figure shows a snapshot of the architecture's activation state when processing the phrase "the red object to the left of the green object" on the scene in Fig. 1a. Dynamic fields are shown as color-coded activation patterns (blue for lowest, red for highest activation), dynamic nodes as circles with activation levels indicated by the intensity of the filling color. For the perceptual field, slices through the three-dimensional activation pattern are shown for the colors green, red, and blue (from top to bottom). Excitatory synaptic connections are denoted by lines with arrow heads, inhibitory connections by lines ending in circles.

representing a template for the spatial term in the phrase (the pattern for 'left of' is visible in Fig. 2; by design, the relational field is always centered on the location of the reference object). This input is excitatory for the relational CoS field, so that the field forms a peak when the relative location of one target candidate matches the region activated by the spatial template. It is inhibitory for the relational CoD field, which forms a peak in the case of a mismatch. The CoS field inhibits the CoD field to prevent it from signaling a mismatch if both a matching and a non-matching target object are present.

In a reverse transformation (green diamond in Fig. 2), a given location in the reference field and a given relative location in the relational CoS field produce a single target location in image coordinates in the *target response field*. Finally, as part of the mechanism for hypothesis testing, the *reference inhibition-of-return (IoR) field* forms a self-sustained peak for any location that has been selected in the reference field, and feeds inhibitory input back to that field.

Processing spatial phrases

As described above, spatial relations are characterized by a reference, a target, and the relation itself. In a concrete spatial phrase these three roles are filled by different concepts, namely, spatial terms and features identifying the target and reference objects (here, we only use color concepts). In the architecture we employ conjunctive coding, in that a pair of dynamic nodes exists for each possible conjunction of role and filler (e.g., 'reference: red' and 'target: red'). Each pair of nodes includes a memory node (blue circles in Fig. 2) and a production node (purple circles). In Fig. 2, each horizontal row of nodes corresponds to one concept (e.g., 'red' or 'right'), while each column of node pairs corresponds to one role (e.g., 'reference').

A spatial phrase is fed into the architecture by activating those memory nodes that correspond to the filler-role conjunctions in the phrase. The memory nodes retain this activation through self-excitation. Each memory node is reciprocally coupled to its corresponding production node, so that active memory nodes pre-activate their production nodes. To become fully active, however, the production nodes need a simultaneous input from an intention node (see below). The production nodes are coupled to different fields by reciprocal, patterned synaptic connections. Color nodes are connected

to different regions of the color intention field, spatial term nodes to different regions in both relational fields. Through these connections, each production node can evoke a specific pattern of activation in the fields (and can conversely be pre-activated by that pattern). Each of these activation patterns is an instantiation of a featural or spatial concept in metric perceptual space. Note that, while the patterned connections have been hand-coded here, they should ultimately be acquired based on neurally realistic learning.

Nodes representing intention and CoS control all aspects of sequentiality. This includes the order in which the stored concepts are invoked, the order in which the roles are filled, and consequently which object is assigned to which role. For each role exists a pair of an *intention node* and a *CoS node* (green and red circles in Fig. 2, respectively). Each intention node drives activation in the corresponding column of production nodes and in specific fields associated with each role (e.g., the reference field). The CoS node in turn receives input from these fields, detects the formation of activation peaks there, and inhibits the associated intention node when the respective role has been filled.

To initiate processing, all intention nodes are simultaneously activated by user input. Sequentiality is enforced by *precondition constraints* in the form of dynamic nodes (black circle marked 'p' in Fig. 2) that inhibit the intention node for one role until the CoS node of another role becomes active. This is employed here to enforce a sequential selection of target and reference object, since both processes rely on the perceptual field.

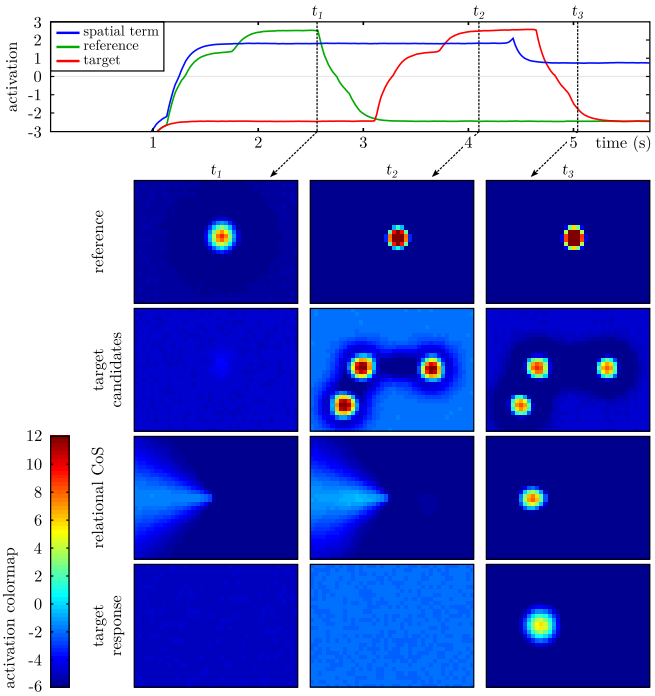

Fig. 3: Evolution of activation patterns for resolving a spatial phrase on the scene in Fig. 1a. Continuous activation time courses are shown for the intention nodes (top), and activation patterns of relevant fields are shown at three selected time steps t_1, t_2, t_3 (bottom). Field activation is color coded (blue for sub-threshold activation, yellow to red for peaks).

Results

In the following, we describe the dynamic processes associated with resolving spatial phrases. All results come from real time numerical solution of the differential equations driven by camera input. The architecture can deal with a variety of differently structured phrases and visual scenes. To simplify visual object recognition, we use scenes with uniformly colored objects on a white background. We illustrate the core capabilities of the architecture using the phrase "the red object to the left of the green object" applied to two visual scenes. While the reference object is uniquely specified by the phrase in the first scene, two identical candidates for the reference object require hypothesis testing in the second example.

Resolving a spatial phrase

We explain how the system resolves the above phrase, with Fig. 1a as visual input. Fig. 3 shows the evolution of activation patterns for this scenario. The spatial phrase is encoded as an activation pattern in the memory nodes, activating the nodes 'reference: green', 'target: red', and 'spatial term: left of'. Processing is then initiated by activating all intention nodes as well as the precondition node. From this point on, the architecture works autonomously.

The intention nodes for reference and spatial term become active, while the target intention node is inhibited by the precondition node. The spatial term intention node boosts all production nodes associated with the role 'spatial term', and thereby activates the node 'spatial term: left of'. This node projects its spatial template into the relational CoS field as a sub-threshold activation pattern (see t_1 in Fig. 3). Analogously, the reference intention node activates the production node 'reference: green'. This node projects into the color intention field, producing a peak at the location corresponding to the color green. This induces a peak in the perceptual field which brings the green object into the foreground. The reference intention node also homogeneously boosts the reference field, which, driven by input from the perceptual field, forms a self-sustained peak at the position of the green object (see snapshots in Fig. 3 at time t_1). This peak means that referent selection is complete, activating the reference CoS node. The CoS node turns off the reference intention node and inhibits the precondition node.

The target intention node can now become active. As it does, it starts to bring red objects into the foreground, whose positions are fed into the target candidates field (see snapshots at time t_2). The positions of the target candidates are transformed and projected into the relational CoS field, where one of them (the top-left one) matches the spatial term 'left of' best and forms a peak (see snapshots at time t_3). This peak is transformed back into image coordinates and fed into the target response field. The correct target object has been located.

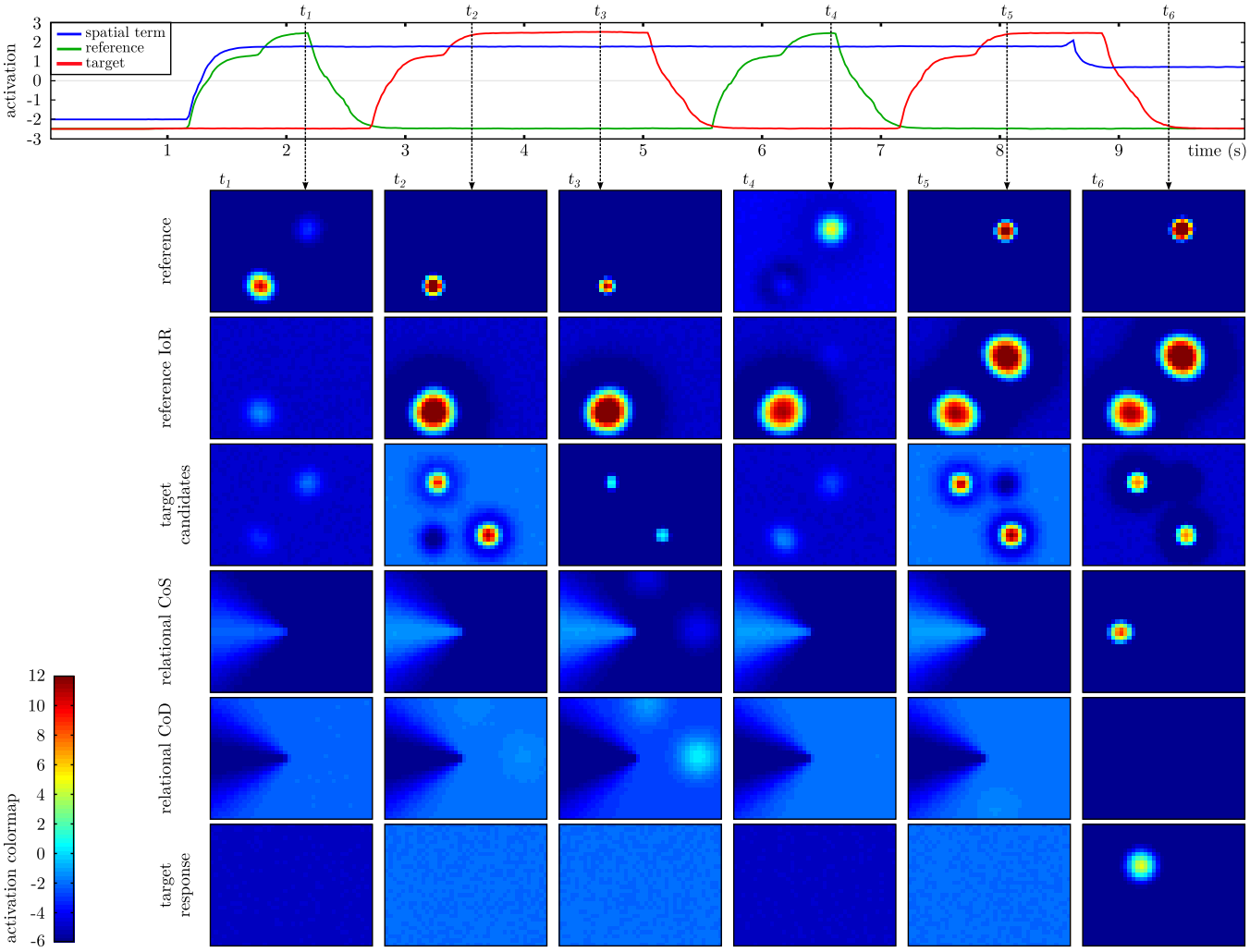

Fig. 4: Evolution of activation patterns for resolving a spatial phrase on the scene in Fig. 1b. Activation patterns are depicted analogously to Fig. 3

Testing multiple hypotheses

We now demonstrate how the architecture can autonomously test hypotheses and discard erroneous ones by resolving the same phrase as above for the scene in Fig. 1b. Activation plots are shown in Fig. 3, with additional fields that are relevant for this more complex scenario.

As in the previous scenario, the spatial template is instantiated and the potential reference objects are brought to the attentional foreground. Faced with two green objects, the reference field autonomously performs a selection decision, forming a single peak for the lower green object (see snapshots at t_1). Its location is also stored in the reference IoR field. Note that the spatial template is visible as inhibitory pattern in the relational CoD field at this time.

In snapshot t_2, the positions of the two red objects have been fed into the target candidates field. Their locations relative to the reference object are determined by the reference frame shift and fed into both relational fields (CoS and CoD).

At t_3, a peak forms in the relational CoD field but not in the relational CoS field, since none of the target candidates is to the left of the chosen reference object. This signals that target selection has failed. The target candidates field and the reference field are inhibited, so that peaks in these field vanish. The target and reference CoS nodes turn off, essentially reactivating the associated intention nodes and restarting the task from the beginning. However, the reference IoR field still retains the memory of the previously selected reference object location, and its inhibitory input prevents this location from being selected again in the reference field.

At t_4, the green object in the top right is established as a new hypothesis for the reference. Subsequently, the architecture identifies the correct target candidate left of that reference. The activation snapshot of the target response field at t_6 shows the position of that selected target in the image.

Discussion

We have shown how a neural dynamic architecture may resolve relational spatial phrases about visual scenes. The specific contribution of this paper is the autonomous control of processing steps and the capability to validate or reject hypotheses about the referents of a relational phrase when only the combination of object description and spatial term uniquely defines the target. Above that, the architecture generalizes to various scenarios, such as answering questions about objects and the spatial relations between them. It is easy to extend the architecture to incorporate additional features beyond color, by adding the associated perceptual, intention, CoS, and CoD fields.

As a process model, the architecture may account for human behavioral data. It currently captures the sequentiality of visual search for target relations, but may provide more specific accounts such as predicting processing time as a function of different forms of cues (Logan, 1994). Similarly, the time course of selection processes in the architecture may be compared to the sequence of attentional shifts that humans perform when they analyze individual relations (Franconeri et al., 2012). Our future work will extend these links to experiment.

The DFT architecture is conceptually related to a number of approaches. In work by Roy, Hsiao, & Mavridis (2004), word meanings are grounded in the physical environment of a robot through a layer between low-level sensory and linguistic representations that implements computational algorithms but is not meant to mirror neural processes. We consider similar computational outcomes, but conceive of the processes as neural activation dynamics directly linked to sensory inputs. Previous proposals within psychology have linked spatial language to processes of visual perception (e.g., Regier & Carlson, 2001). We have moved beyond those proposals by providing specific process accounts based on the well-established neural dynamic mechanisms of detection, selection, and working memory. Moreover, we have accounted in a principled way for how the succession of processing steps is generated autonomously.

At a more general level, the architecture resonates with the idea that relational concepts may be embedded in modal neural processes (Barsalou, 1999). All relational operations occur within neural fields, close in format to how perceptual information is represented. The discrete activation nodes are more akin to amodal representations but primarily organize the processing in time. Mapping non-spatial concepts onto spatial representations may provide a route toward extending the ideas of this model to general cognition (e.g., Knauff, 2013). The present work is only a first step toward a neurally grounded account of higher cognition.

Acknowledgments

The authors acknowledge the financial support of the European Union Seventh Framework Programme FP7-ICT-2009-6 under Grant Agreement no. 270247—NeuralDynamics.

References

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(04), 577–660.

Burigo, M., & Knoeferle, P. (2011). Visual attention during spatial language comprehension: Is a referential linking hypothesis enough? In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Springer.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1-2), 25–62.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–27.

Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. Cambridge, MA: MIT Press.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(6).

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1015–1036.

Logan, G. D., & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (pp. 493–529). Cambridge, MA: MIT Press.

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

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2457–2464).

Roth, J. C., & Franconeri, S. L. (2012). Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in Psychology*, *3*(November), 464.

Roy, D., Hsiao, K.-Y., & Mavridis, N. (2004). Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics*, *34*(3), 1374–83.

Sandamirskaya, Y., & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179.

Schneegans, S., & Schöner, G. (2008). Dynamic Field Theory as a framework for understanding embodied cognition. In P. Calvo & T. Gomila (Eds.), *Handbook of Cognitive Science: An Embodied Approach* (pp. 241–271). Amsterdam, Netherlands: Elsevier.