Cognitive Neurodynamics 3(4), 2009
http://dx.doi.org/10.1007/s11571-009-9096-y

Swing it to the Left, Swing it to the Right: Enacting Flexible Spatial Language Using a Neurodynamic Framework

John Lipinski · Yulia Sandamirskaya · Gregor Schöner

Received: 8 June 2009 / Revised version: 27 August 2009 / Accepted: 2 September 2009

Abstract Research is continually expanding the empirical and theoretical picture of embodiment and dynamics in language. To date, however, a formalized neural dynamic framework for embodied linguistic processes has yet to emerge. To advance embodied theories of language, the present work develops a formalized neural dynamic framework of spatial language that explicitly integrates linguistic processes and dynamic sensory-motor systems. We then implement and test our spatial language architecture on a robotic platform continuously linked to real-time camera input. In a suite of tasks using everyday objects we demonstrate the framework's capacity for both contextually-dependent behavioral flexibility and the seamless integration of spatial, non-spatial, and symbolic representations. To our knowledge this is the first unified, neurally-grounded architecture integrating these processes and behaviors.

Keywords: dynamical systems, spatial cognition, spatial language, neural fields

Institut für Neuroinformatik
Ruhr-Universität Bochum
Bochum, Germany
Tel.: +49-234-3224201
Fax: +49-234-3214209
E-mail: 2johnlipinski@gmail.com

1 Introduction

Theories of cognition are often dissociated from the real-time generation of behaviors. This is particularly true in the domain of language, where theoretical treatments tend to emphasize highly abstracted concepts and symbolic representations (e.g. [47]).

Recently, however, attention has shifted to how language is produced and experienced by real bodies in the real world[19]. Work from beim Graben and colleagues [41,40], for example, shows that non-linear dynamical systems approaches can enhance our understanding of syntactic processing within neural systems. A recent review by Elman [25] further highlights how dynamical models (e.g. simple recurrent networks) can shed light on language processing, particularly contextual dependencies in grammar and word learning. The role of context is also increasingly prominent in recent models of language development, revealing how language processing dynamics shape infant categorization [37], lexical organization [54] and task-specific noun generalization behaviors [70].

This attention to the contextually rich, coordinated dynamics of language is part of a growing view that linguistic processes are embedded within a broader embodied, dynamic

system intimately linked to the physical world [34,5,81]. The evidence supporting the role for the body and real-time dynamics in language is broad and includes motion-dependent action word processing [35], the activation of motor circuits when listening to action-related sentences [88], signatures of continuous dynamic processes in spoken word recognition and semantic categorization [82], and the synchronization of speech and gesture [61,62]. Yet, despite this expanding empirical and theoretical picture, a formalized theoretical framework for embodied linguistic processes has yet to emerge.

Spatial language provides a useful entry point for developing such a framework because it is an elementary link between integrative linguistic processes and the embodied, dynamic sensory-motor systems that fluidly operate in the spatial world. Given these embodied roots, a viable spatial language framework must specify how differing behaviors (e.g. language production and language-guided action) can emerge from the same system, how non-spatial object features (e.g. color) can be integrated with spatial information, and how linguistic symbols can be tied to the continuous sensory-motor representations.

Some spatial language theories to date have touched on related embodiment issues. Recent modeling work, for example [21,13], accounts for empirical results showing that functional relations between objects influence spatial language behavior. Regier and Carlson [69] have also provided important insights into the complex contributions of attention and landmark shape. However, these models do not generate flexible behaviors in real-time nor do they provide transparent accounts of the representational integration supporting this flexibility. This represents a substantial theoretical gap. We contend that a process-based account of spatial scene representations and behaviors derived from these representations is required to address this gap.

Recent work suggests that a systems-level neural-dynamic approach to human cognition can provide the conceptual foundation for such a process-based account. At the broadest level, this perspective argues that behaviors unfold in real-time from the continuously coupled interplay between neural-dynamic decision processes, the sensory-motor system, and the feature-rich environment in which bodies are embedded [84,91,9,90,75]. These approaches have established strong contact with observable human behaviors across a variety of contexts, including saccadic eye movements [94], visual discrimination and visual working memory [78, 48,49], spatial working memory development [77,76], and infant reaching errors [89]. These empirical ties suggest that complex, integrative spatial language behaviors may be similarly described in neural dynamic terms. To advance embodied theories of language, the present work therefore seeks to develop and test a formalized neural-dynamic architecture of spatial language.

To this end, we first discuss three key characteristics of embodied spatial language, namely behavioral flexibility, the integration of spatial and non-spatial features, and the integration of symbolic and continuous representations. Next, we briefly outline three principles of a neural dynamic system that collectively address these aspects: gradedness, autonomy, and stability. With this conceptual background we introduce the Dynamic Field Theory (DFT), a neurally-based theoretical language that incorporates activation profiles defined over continuous dimensions and emphasizes attractor states and their instabilities [75, 79]. The DFT is the foundation of our neural-dynamic architecture.

The gradedness, stability and autonomy of the DFT framework allow one to couple the cognitive architecture to the sensory-motor system. To demonstrate this capacity we implement and test our spatial language architecture on a robotic platform. Importantly, we use the low-level sensory input provided by the robot's camera. Thus, our model deals with the problem of extracting the categorical, cognitive information from the low-level sensory in-

put through the system dynamics, not through the preprocessing of the visual input in an ungrounded, neurally implausible way. Models which do not directly link cognitive behavior to lower-level perceptual dynamics risk sidestepping this difficult issue. Our explicit connection to behavior through the robot provides a key demonstration of sufficiency of our neural-dynamic approach and a heuristic for understanding how spatial communication emerges from lower-level sensory dynamics.

In a suite of varying spatial-language tasks using everyday objects we demonstrate the framework's capacity for both contextually-dependent behavioral flexibility and the seamless integration of spatial, non-spatial, and categorical representations. In doing so, we draw particular attention to the time course of these behaviors, thereby revealing the neuro-dynamic roots of representational integration and behavioral flexibility within our spatial language system. To our knowledge, this is the first unified neurally-grounded architecture that integrates these processes and behaviors. As such, our system represents a step towards the development of a more comprehensive, neural-dynamic model of human spatial language and embodied language processes more generally.

1.1 Flexibility and Integration in Spatial Language

The neural dynamics of any language behavior are immensely complex and multifaceted. To develop a conceptually manageable framework, we focus on three core aspects of spatial language that arise from its embodied roots, namely behavioral flexibility, the integration of spatial and non-spatial representations, and the integration of categorical and continuous representations. The present section considers each in turn.

The power of the spatial language system is revealed in its broad behavioral range, from following directions [24] and creating mental models [87] to telling stories [53] and coordinating joint attention and action [4,50,74]. Even within a single, highly constrained environment such as a shared tabletop workspace, spatial language exhibits an impressive degree of flexibility.

Consider, for example, a cluttered office desk in which a cup of coffee sits to the right of a laptop computer. Given this context, the human spatial language system can freely generate descriptions of object-centered relations in the scene (i.e. spatial descriptions that select another object as a reference point). Thus, if one asks "Where is the green coffee cup in relation to the laptop?" then a person with knowledge of the scene can easily answer "To the right." On the other hand, if one asks "What is to the right of the laptop?" one viewing or remembering the scene could instead respond "The green coffee cup." The production behavior in both these instances, of course, also assumes the complementary capacity to comprehend the questions. Moreover, both spatial language production and comprehension flexibly process different reference objects and spatial terms across highly variable visual scenes – people can use spatial language to describe just about anything. Behavioral flexibility is thus part and parcel of functional spatial communication.

Our second critical aspect is the integration of the fine-grained, metric sensory-motor representations [48,39] with the categorical, linguistic representations rooted in the language faculty [44,66]. To successfully index items in the world one must map the symbols of language to the dynamic representational states of perception. Referring to the coffee cup, for example, assumes the ability to link information in the visual system to words like "green", "cup", "laptop", and "right". This link is of course functionally bidirectional, enabling us to produce language about the visible world and map the words we hear onto a visual scene. Moreover, because spatial language can be used to guide others' behaviors, this representational integration also extends into the motor processes controlling behavior. Considered together, these aspects highlight the need to ground

language in the neural dynamics underlying scene representations in a manner that permits flexible manipulation of the symbolic units (for related discussion see also [36,5,65,34,11,43]).

Our third point of focus is the integration of spatial and non-spatial features. Consider again our description of the green coffee cup that sits to the right of the laptop. In this case, the individual must process both the explicit spatial term "right" and the non-spatial descriptor "green" to identify and use the landmark. This link between categorical spatial relations (e.g. right) grounded in metric space and non-spatial perceptual features (e.g. green) enables one to reference landmarks within the visible (or remembered) environment using non-spatial object characteristics such as color, texture, or size. The ability to integrate different features is thus central to generating and comprehending spatial descriptions.

1.2 Embodied Cognition: Supporting Neural Dynamic Concepts

To this point, we have identified three critical aspects of spatial language that a viable neural-dynamic approach must address. As we previously noted, extant spatial language models have addressed a number of important dimensions that speak to the embodiment of spatial language including attention, landmark shape, and functional features (e.g. [21,13,69]). Nonetheless, no model to date has brought the detailed aspects of behavioral flexibility and representational integration together within a single framework.

The limits of current spatial language theories arise from the failure to provide a neurally-grounded account of real-time spatial language behaviors and their roots in spatial scene representations. Consequently, current theories typically overlook some questions fundamental to understanding representationally complex, embodied spatial communication. For example, how do neurally-grounded scene representations develop over time on the basis of sensory information? How do the dynamic processes supporting scene representation shape the time course of spatial language behaviors? How do context-specific inputs like spatial terms and visible objects dynamically structure the integration of the multiple components supporting behavioral flexibility? Developing a neurally grounded, formalized framework is a key step to answering such questions.

What are the concepts underlying such a neural framework? The first such concept is autonomy. Autonomy means that neural processes unfold continuously in time on the basis of both past and present neural states and past and present sensory information. As a result, the autonomous cognitive systems are sensitive to input, but not purely input-driven [71].

Autonomy is critical for a cognitive system because it provides the basis for structuring behavior in a context-dependent manner. To be effectively adaptive, cognitive systems must be able to smoothly flow from decision to decision and action to action in accord with both the current environment and the behavioral context [75]. Autonomy makes this flexible and continuous integration of goals, decisions, and actions within an embodied system possible. Without autonomy, a neural dynamic system would not be able to modulate the multi-dimensional integration supporting this flexibility and would instead more closely approximate input-compute-output processes or stimulus-response associations.

Representational gradedness is the second neural concept central to describing cognitive processes grounded in the sensory-motor system. A graded representation of a behavior or percept is defined over one or several continuous feature dimensions which constitute the behavior or percept linked to the motors or sensors. Within an autonomous neural dynamic system, the computations taking place over these graded representations may be described using concepts from non-linear dynamical systems [27]. The neural dynamics of movement preparation, for example, may be characterized according to non-linear signatures emerging over the continuous dimensions of reach-

ing amplitude. Low-level visual processing, on the other hand, may be described by the neural dynamics of spatial and non-spatial metric features (e.g. color) available in the visible scene. Importantly, metric features have also been shown to shape spatial language behaviors [56,69]. As a result, such graded, metric features, which are critical to sensory-motor dynamics [7,8,27] and non-linguistic decision processes [48,49], may also be used to probe the neural dynamics of spatial language.

The successful integration of graded sensory-motor representations with the spatial language system depends on the notion of stability, a core principle of dynamical systems thinking and the third neural concept we emphasize. Stability is the capacity of a dynamical system to resist change. It thus plays a central role in the neural dynamics of cognition because it provides for consistent behavior in the face of neural or environmental noise. In the absence of stability, graded representations grounded in the sensory-motor system would be subject to continual shifts arising from inherently noisy neural states. Stability is therefore a prerequisite for the grounding of sustained cognitive behavior on neural-dynamic states linked to sensory data [75].

Observe, however, that to be adaptive, autonomous dynamical systems must also be able to destabilize and form new stable states as the contexts and behaviors demand. This balance between stability and instability is fundamental to behavioral flexibility and is the prime challenge for formalized theories of autonomous embodied cognition.

The Dynamic Field Theory incorporates each of these concepts and therefore provides the representational foundation of our proposed framework. We introduce this theory in the following section.

2 Methods

2.1 Dynamical Field Theory (DFT)

Dynamical Field Theory is a neural-dynamic approach to embodied cognition in which cognitive states are represented as distributions of neural activity defined over metric dimensions. These dimensions may represent perceptual features (e.g., retinal location, color, orientation), movement parameters (e.g., heading direction, end-effector velocity) or more abstract parameters (e.g., location relative to an object, visual attributes of objects like shape or size). These metric spaces are continuous, representing the space of possible percepts, actions, or objects and scenes. They are endowed with a natural metric which represents perceptual or motor similarity.

Spatially continuous neural networks, or neural fields, were originally introduced as approximate descriptions of cortical and thalamic neuroanatomy based on the spatial homogeneity of cortex along its surface [96]. The principle of topographic mapping of feature spaces onto cortical surfaces [51] has been evoked to extend the notion of neural fields to dimensions beyond the neuroanatomical ones [93] (see [6] for critical discussion). More recently, however, the notion of Distributions of Population Activation shows how neural fields may describe neural representations of metric dimensions independently of neuroanatomy [26,8,18]. Each neuron contributes its tuning curve to the representation of a feature dimension, weighted with its current firing rate. As a result, neurons are not localized within the neural fields, but distributed according to the specificity of their response. A single, localized peak in such a Distribution of Population Activation represents a specific value of the metric dimension, but potentially involves broad populations of neurons that may be spatially distributed. The population vector reflects both the specified metric value and the total amount of activation [33].

Neural fields are recurrent neural networks whose temporal evolution is described by iteration equations. In continuous form, these take the form of dynamical systems. The mathematics of dynamical neural fields was first analyzed by Amari [2] and much modeling has since built on the original Amari framework [73] which we briefly review here. The activity distribution of a neural field $u(x,t)$, defined over a continuous metrical space $\mathbf{X}$, $x \in \mathbf{X}$, evolves in time according to

$$\tau \dot{u}(x,t) = -u(x,t) + h + I(x,t) + \int f(u(x',t))\omega(x-x')dx'. \quad (1)$$

The rate of change, $\dot{u}(x,t)$, of the field's activation at a time, t, and a field site, x, is proportional to the negative of the current activation level, $u(x,t)$. This provides the fundamental stabilization mechanism. Added to this are a negative resting level $h < 0$, inputs from sources outside the field, $I(x,t)$, and inputs from other sites, x', of the same neural field. This last term represents neural interaction and is characterized by excitatory coupling among field sites that are close to each other and inhibitory coupling across larger distances:

$$\omega(\Delta x) = c_{exc} \exp[\frac{(\Delta x)^2}{\sigma_{exc}^2}] - c_{inh} \exp[\frac{(\Delta x)^2}{\sigma_{inh}^2}] \quad (2)$$

(where $\Delta x = x - x'$ and $\sigma_{exc} < \sigma_{inh}$).

Only sites with sufficient activation contribute to this lateral interaction as described by the sigmoidal non-linearity:

$$f(u(x,t)) = \frac{1}{1+e^{-\beta u(x,t)}}, \quad \beta > 1 \quad (3)$$

Thus, while activation stays below the threshold of this sigmoid (defined as the zero level of activation), interaction plays a minor role in the evolution of the field, which is dominated by input. This is true for sufficiently weak external inputs $I(x,t) < h$. With increasing external input, activation of the field, $u(x,t)$, surpasses the threshold and interaction begins to engage at locations at which $f(u(x,t)) > 0$. This induces a bifurcation in the field dynamics, the so-called detection instability [75]. Beyond this instability, localized peaks of activation are self-stabilized: activation is stabilized against decay by local excitatory interaction and stabilized against diffusive spread by global inhibitory interaction. The resulting localized peak of activation is a unit of representation in the Dynamic Field Theory approach to cognition. When multiple fields are coupled, as will be the case in the model developed here, the detection instability is critical also for the propagation of activation from one field to another. The extent to which such peaks are sensitive to further changes of input depends on the strength of interaction. Fields with strong interaction support self-sustaining peaks which maintain activation after the complete removal of the external input that initially induced them. Such peaks are robust to new distractor input and comprise a form of working memory (see e.g. [76,49,78,80]).

From spatially continuous fields, categorical states may emerge. This is based on the same mechanism as the detection instability which may amplify small inhomogeneities in the field into macroscopic peak states [75]. Assume, for instance, that a few locations in a field are frequently activated. Generic learning mechanisms may change the neural dynamics such that these field locations are more excitable than other, less frequently activated locations. If a broad input is now applied to the field, one of the more excitable field locations may be the first to be pushed through the threshold at which interaction engages. A full-fledged self-stabilized peak will develop at that location, which then prevents additional peaks from being generated at other locations through inhibitory interaction. This peak reflects categorical behavior, because the field location depends on the learned inhomogeneities in the field, not on the spatial structure of the inducing input.

In this paper, we will not address the learning mechanisms through which learned inho-

mogeneities in the field arise. Instead, we will use an effective dynamical description of such categorical behavior by introducing discrete dynamical neurons with self-excitatory interaction, which represent the activation at excitable field sites. Given sufficient input to such neurons, a detection decision is made at which the neuron switches to an activated state. This state can stabilize itself against weaker input in a bistable regime. Conversely, a discrete neuron may provide localized input to an activation field exactly like a localized peak of activation in a field does. Dynamical Field Theory thus provides a framework for integrating metrically continuous and metrically discrete categorical representations.

2.2 The spatial language architecture

Spatial language is a complex behavior that draws on numerous cognitive processes including vision, spatial cognition, and language. Spatial language behaviors therefore depend on numerous cortical and subcortical regions. Comprehending or producing spatial language about a visual scene, for example, not only involves a neural scene representation that emerges from the retinal image but also the integration of long-term memory about objects and their features and the neural representation of spatial semantic terms (e.g. right, above, etc). Critically, these semantics must be applied to the current scene and they are often aligned with a reference object [52]. Not surprisingly, the neural populations accomplishing these various functions are widely distributed over the cortex with V1-MT processing visual features [42], the parietal cortex supporting spatial representation and reference frame transformations [3,20,22], and the frontal, inferotemporal, and the temporal-occipital-parietal junction regions supporting spatial language [23,92].

Our model is similarly distributed and contains several interconnected modules each maintaining a unique functionality that affects the dynamics of the other modules. The Feature-space fields (Fig. 1A) are driven by the visual input and represent the locations and features of objects. The Reference field (Fig. 1B) represents the reference object location, the point relative to which spatial terms are defined. The Spatial semantic templates (Fig. 1C) express the semantics of the spatial terms. These templates are aligned with the location of the reference object by the "shift" mechanism (Fig. 1D) and then integrated with the visual, feature-based object representations in the Spatial semantic neural fields (Fig. 1E). The language terms specifying features ("red", "green", "blue") or spatial terms ("left", "right", "above", "below") are represented by bi-stable dynamical nodes, which are interconnected with the neural fields of the model. These connections express the semantic meaning of the particular term.

Although our network is distributed, each functional component is based on the same dynamical neural field principles. This creates a functionally and theoretically coherent spatial language architecture that not only respects core neural principles but which can also be linked to real-world sensory information and behavior.

To provide the most rigorous test of our embodied approach, we implement the model on a robot that is equipped with a vision system. In doing so, we directly link the sensory and motor properties of the robot to the internal neural dynamics of the system, bringing this robotic implementation in line with known principles of the human nervous system.

The sections below detail these functional modules and the robotic implementation.

2.2.1 Representing locations and colors

When using spatial language, people often refer to spatial relations between objects. In "The toaster is to the right of the sink", for example, the toaster's location is defined relative to the sink. To either produce or act on this spatial information, both the target object (the toaster) and the reference object (the sink) must be identified in the visual scene.

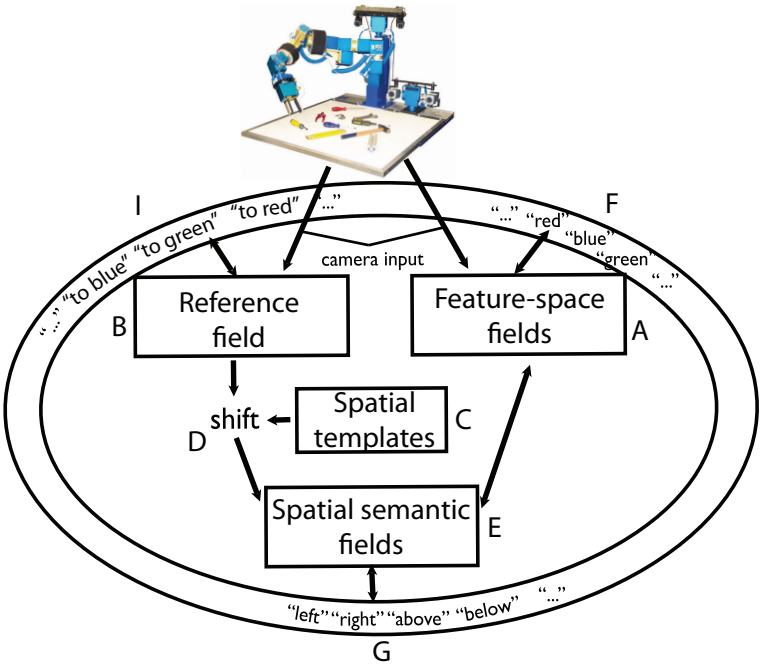


Fig. 1 Overview of the neural-dynamic spatial language architecture

The image of the visual scene on our retina is first processed by the early visual regions of cortex, where the perceptual features are extracted and retinotopic feature maps are built [30, 42]. In the DFT, these feature maps are modeled as a set of feature-space fields. Each site of these dynamical fields is sensitive to a particular value of a visual feature at a certain retinotopic location.

In the present work we focus on color as the visual feature. Color – as with other local features such as orientation and texture – is known to contribute to representations in early visual processing [31] (see also [28] for a comprehensive DFT approach to multi-dimensional object representation). As a low-level feature with an underlying metric, color can be also easily mapped onto a continuous dimension of a neural field.

The color-space neural field is a three-dimensional dynamical field $F(x, y, c; t)$, each site of which responds to a color, c, at a certain location on the retina, (x, y). The activity distribution in this field thus represents the color distribution in the visual scene. A localized blob of a certain color in the scene can potentially give rise to an activation peak in the color-space field. Such a peak is a dynamic object representation grounded in the object's graded location and color distribution.

In our implementation, the color dimension is resolved sparsely, because we require only a few colors to represent the objects used in our demonstrations. The three-dimensional dynamical field is therefore implemented as a stack of six two-dimensional dynamical fields. Each of these color-space fields is a two-dimensional field whose sites respond to the spatial position of a particular color. These fields are globally inhibitory such that an activation peak within one field leads to a uniform inhibition of the remaining color-space fields.

The visual input to the color-space fields is provided by a robotic camera. The camera image here plays a role similar to that of retinal images in human cognition. The process

of extraction of feature maps from the retina images is substituted by a color extraction algorithm. The result is a distribution of colors defined over the space of the image plane. These distributions correspond roughly to the retinotopic feature maps found in early visual processing [30]. In particular, the color is extracted from the camera image as the hue of each pixel in the hue-saturation-value (HSV) color space. This hue value is binned according to one of six equidistant hue ranges (representing the basic colors red, orange, yellow, green, blue, and violet) and provides input to the corresponding color-space field. The input into the color-space field location matches the pixel's image location. The pixel's intensity (value) determines the strength of this input.

Fig. 2 illustrates this process. The visual scene here consists of three objects: a green tube, a blue wire-roll and a red plastic apple (Fig. 2A). They provide inputs to the color-space fields (mainly to "green", "red", and "blue" fields respectively) at positions corresponding to the locations of the objects in the image (see Fig. 2B). These localized input activations to the color-space fields are subthreshold. This means that when the input is summed with the negative resting levels, the activation at the specified field sites remains negative. Thus, the fields produce no output and no activation is propagated to other sites in the fields or to other parts of the architecture. When a localized activation surpasses threshold, however, output is produced that is then passed to other field sites and other parts of the model. This activation plays a key role in the dynamic structuring of activity in these other elements and, ultimately, the generation of task-specific linguistic and motor behaviors (see below).

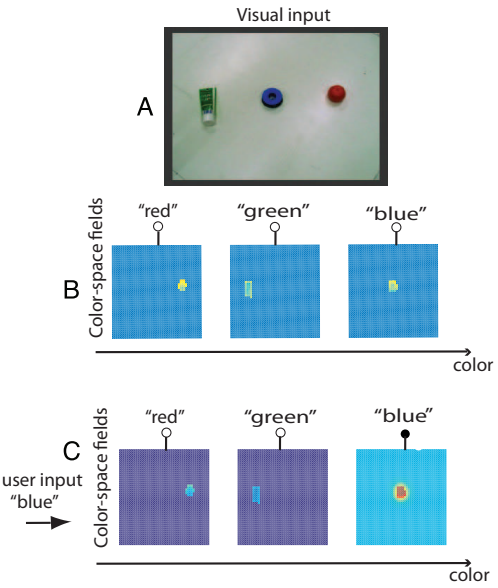


Fig. 2 Panel A: The visual scene containing three objects – green tube of creme, blue tape-roll and a red plastic apple – provides input to the color-space fields (**Panel B**; three fields shown here). **Panel C**: specification of the "blue" color term activates the "blue" color-term node, raising the resting level of the "blue" color-space field. The peak of positive activation in the "blue" color-space field represents the location of the blue object in the image.

2.2.2 Representing color terms

When people refer to objects in the world, they link discrete linguistic representations to the graded, metric features of the visible world. The exact nature of these connections in cortex has yet to be identified, however (although for work in this direction see [60,38]). In our model, we represent language terms by simple bi-stable dynamical nodes within a winner-take-all network of competing nodes.

Each discrete node is reciprocally linked to one of the color-space fields. The discretization of the three-dimensional color-space field thus provides the linguistic categorization along the color dimension. Such linguistic mappings between the discrete linguistic and the underlying continuous feature maps are hypothesized to emerge from experience over development. Because we are aware of the categorization properties of the neural fields [95] and emergence of the color categories is not of interest for this paper, we allow this simplification here.

Color-term nodes may become active through external linguistic input. The color-term node's activation is further propagated along its link to the color-space field. Fig. 2C illustrates the

linguistic boost effect, in which the user-specified linguistic input "blue" activates the "blue" color-term node, thus raising the resting level of the "blue" color-space field and pushing the activation there beyond the detection threshold. When this threshold is surpassed, the active sites in the field engage in lateral interactions. This induces a localized activation peak whose location corresponds to that of the target object in the camera image.

In addition to linguistic input, a peak in a color-space field can also be driven by positive activation coming from other parts of the cognitive architecture. Because the color-term nodes are reciprocally linked to the color-space field, such a peak would increase the activation of the linked color-term node and trigger the generation of a descriptive color term. The color-term nodes thus provide the means of generating a specific color term description as well as processing linguistic input from the user.

2.2.3 Reference field

To describe a target object location by reference to another object (e.g. "The toaster is to the right of the sink"), the reference object location must also be represented in a feature-dependent manner. The reference field (Fig. 1B) serves this role in our framework. In our implementation, the reference field is a two-dimensional neural field (Fig. 3C) that receives visual input (Fig. 3A). This input is modulated by the reference color-term node, which specifies the color of the reference object. The color information is then extracted from the camera image in a manner similar to that of the color-space fields (section 2.2.1); only those pixels with the color specified by the reference color-term node serve as input to the reference field.

The reference field is always in the "detection" mode. This means that an object of the specified color always induces an activation peak in this field. This peak, which represents the location of the reference object in the image, is stabilized by the interactions in the field. Nonetheless, it is also updatable if the reference object moves.

2.2.4 Spatial semantic templates

Spatial language terms typically represent prototypical regions ([45,58], although see also [21]). Thus, saying "left" usually highlights the same part of the visual scene for English speakers. These semantically specified spatial regions, or "templates" [58], may be described by weight matrices in which regions corresponding to prototypical instances of the term have higher weight strengths. Regions which provide a poorer fit with the spatial term, on the other hand, have lower weights.

In the present architecture, the precise connection weights between the four spatial term nodes ("left", "right", "below", and "above") and the spatial fields are based on a neurally-inspired approach to English spatial semantic representation [64]. These connection weights are defined by Gaussian distributions in polar coordinates (see Eq. (4) and parameter values in Appendix). When viewed in Cartesian coordinates as applied here, they take on a teardrop shape (see Fig. 3B):

$$M = \exp\left[-\frac{(\rho-\rho_0)^2}{2\sigma_\rho^2}\right]\exp\left[-\frac{(\theta-\theta_0)^2}{2\sigma_\theta^2}\right] \quad (4)$$

2.2.5 Spatial semantics alignment

As previously mentioned, spatial terms are often used in conjunction with a reference object (see Reference field in section 2.2.3). Consequently, spatial semantic templates must be aligned with the reference object location. However, objects are initially represented in the retinotopic rather than object-centered reference frame. The spatial templates must thus be dynamically coupled with the space of the visual scene to permit the flexible use of spatial descriptions that are anchored in the world.

Although the exact neural mechanism of this reference frame transformation process has

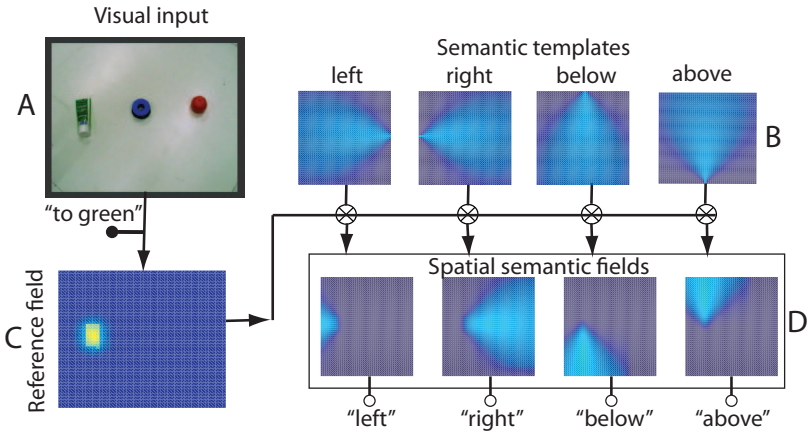


Fig. 3 Reference field and spatial semantics alignment. **Panel A** shows the camera image containing three objects (green toothpaste tube, blue wire roll, red plastic apple). **Panel B** shows the spatial distribution of the weight strengths for each of the four spatial semantic terms (lighter blue regions indicate greater weight). **Panel C** shows activation in the two-dimensional reference field. The activation peak (yellow blob) corresponds to the green object location identified as the referent in this example. **Panel D** depicts the spatial semantic fields with input from the semantic templates (Panel B) aligned with the reference object location (i.e. the light blue region in the "right" spatial semantic field represents region to the right of the green reference object).

yet to be identified, different solutions are possible [67,55]. In the present work we solve this problem through a spatial template "shift" mechanism (Fig. 1D) which aligns the semantic templates with the position of the reference object. The semantic templates are only allowed to contribute to the spatial semantic field dynamics (see below) after this has occurred.

We implement this "shift" or "alignment" of spatial semantics as a convolution of the output of the reference field, which holds the reference object position, with the semantic template functions. Because the reference object is represented by a localized activation pattern, the convolution centers semantic weights on the reference object location. The shift of the semantic weights can thus be viewed as a modulation of the synaptic connection strength between a spatial term node and the spatial semantic field according to the activation in the reference field. Fig. 3D shows an example of this spatial semantic alignment in which the semantic weights are centered on the location of the green reference object.

2.2.6 Spatial semantic fields

For the system to process spatial language about the visual scene, spatial information about the target object and the aligned spatial templates must be integrated. In our model, the spatial semantic fields provide this function (Fig. 1E) Spatial semantic fields are neural arrays with weak dynamical field interactions (see parameter values in the Attachment). Each spatial semantic field is associated with one spatial semantic template. Each spatial semantic field therefore represents a single spatial relation ("left", "right", "above", or "below" in the present implementation), Fig. 3D.

The spatial semantic fields each receive activation from the color-space fields which specify the target location. By blending this target location information from the color-space fields with the aligned semantic weights, the spatial semantic fields integrate the target and spatial term information, thereby linking spatial term knowledge to the visual scene.

In addition, each spatial semantic field is also reciprocally linked to a categorical spatial-term node, analogous to the color-term nodes

("left", "right", "above", or "below" node; see Fig. 1G). If the activation within a spatial semantic field is sufficient, it will trigger the activation of the linked node, signaling the selection of one of the four represented spatial categories. In addition, this node can also receive external linguistic input. This linguistic activation of a spatial term boosts the activation of the linked field and can thus contribute to the dynamics of the system.

2.2.7 Linguistic input and motor output

To communicate with the robot, we use a graphical user interface (GUI), not speech input. Nevertheless, the implemented interface does incorporate some properties of the real-world communication. In particular, the order of the linguistic inputs and the timing interval between them are arbitrary rather than fixed. Moreover, the model integrates these GUI inputs continuously in time, just as the human nervous system continuously integrates linguistic inputs. The timing of the input and its contribution to the internal dynamics are therefore flexibly determined by the user. For this reason, sustaining this characteristic flexibility of natural language is a non-trivial property of the spatial language framework.

To generate a motor behavior, we implemented a dynamics controlling the camera-head configuration (pan and tilt; see Appendix). Attractor dynamics are known to be a viable model for many human motor behaviors [27]. The dynamical system implemented here has an attractor that is effectively set by the localized activation peak at the target object location (as represented in the color-space fields), forcing the robot to turn the camera head and center the attended object in the corresponding field. This coupling of the spatial language architecture to motor behavior further highlights the power of the neural dynamic framework to integrate higher- and lower-level processes within a single system.

In generating this motor behavior, it is important to note that such sensor movements change the spatial relations between the objects out in the world and the robot's sensory surface (image plane). Yet, these very spatial relations continuously structure the camera movement dynamics. Consequently, moving the camera potentially disrupts the visual inputs on which the contextually-adaptive camera movements depend. For this reason, camera movements in our dynamically integrated system provide a rigorous test of the model's stability properties.

3 Results: Demonstrations on a Robotic Platform

Our goal is to model the neural dynamic processes supporting flexible spatial language behaviors within a unified system. Such behaviors include generating a spatial description of an object location (e.g. "The apple is to the right of the toaster") from visual input and localizing objects in a scene based on a linguistic description. Because our robotic implementation links a formalized neural-dynamic model with visual input, we can test the real-time behavioral flexibility of our model through linguistically and visually varied dialogues. We here detail our model's performance in five such scenarios.

Each demonstration combines real-world visual input with user-specified linguistic input provided through the GUI. The robot's task is to either (a) select a descriptive color or spatial term that matches the described target object or (b) build a peak at the described target location. Thus, in Demonstration 1a, for example, we ask "Where is the blue object relative to the green one?" and the robot must choose the correct spatial term. In Demonstration 4b, on the other hand, we ask "Where is the red object to the left of the blue object?" and the robot must select the correct object by building a peak at the correct location and centering that object in the visual image.

In providing the linguistic input through the GUI, it is important to note that appropriate selection decisions do not in any way de-

pend on the sequence or the timing intervals in this input. Indeed, as we show below, the autonomous neural dynamics of our system are at once continuously sensitive to new linguistic inputs but nonetheless behaviorally robust with respect to the fine-grained timing details – getting the right answer does not depend on careful input timing. In this vein, we further observe that the localist nodes activated by these linguistic inputs can be used in different ways in different tasks. In some instances, node activation drives activity in a continuous field. In others, node activation represents a decision driven by the internal neural dynamics. Because these nodes can be flexibly operated upon, they provide key symbolic functionality.

Importantly, the flexibility in timing of the human-robot interaction is achieved by the attractor dynamics. Being in an attractor state, the system can sustain variable time intervals between user actions. Keeping in mind that the real-time behaviors and interaction with the user are central in our work, we measure time in our demonstrations in physical units (seconds) rather than the more conventional simulation time-steps. Because the system relaxes to an attractor state rapidly – as guaranteed by the choice of the time-constant of the dynamics $\tau \approx 2.5ms$ – the timing of the relevant events in the system is more sensitive to the real-world processes than to the computational power of the computer hardware. To maintain consistency across the demonstrations, we kept this notation even when showing the cascade of instabilities leading to a single decision in the framework when the user input and the perceptual input did not change.

3.1 Demonstrations 1a and 1b: The neural dynamics of "Where" and "What"

One basic function of spatial language is to describe *where* an object is. Another basic function is to learn about *what* object occupies some described space. Our architecture dynamically integrates spatial and feature-specific linguistic input through metric visual information, giving rise to two basic interactive pathways: a "Where" pathway and a "What" pathway. Our model therefore directly addresses these two basic functions.

When probing the "Where" pathway in Demonstration 1a, the user specifies the target and reference objects and the robot provides a descriptive spatial term response (i.e. a "Where" response). For instance if the user specifies that the target object is blue and the reference object is green, this would be analogous to asking "Where is the blue object relative to the green one?" and expecting a descriptive spatial term in response. When probing the "What" pathway in Demonstration 1b, on the other hand, the user specifies the spatial term and the reference object and the robot selects a color term that describes the target object (i.e. a "What" response). For example, if the user specifies that the reference object is blue and the target is to the right of this referent, this would be analogous to asking "What is the color of the object to the right of the blue one?" and expecting a descriptive target color response. Note that the specification of the reference object's color is obligatory in these and the following scenarios (although we could use the default reference point in the center of the working space otherwise).

We examine the dynamics of these two scenarios below by combining the linguistic input with a visual scene of three objects – a green tube, a blue wire-roll and a red stack of blocks – approximately aligned horizontally (see Fig. 4A and Fig. 6A). By integrating linguistic input, visual input, and both spatial and non-spatial feature values in two different tasks, these demonstrations provide the conceptual building blocks for the additional tests that follow. To our knowledge, they also represent the first evidence of behavioral flexibility within a unified, neurally-grounded spatial language model.

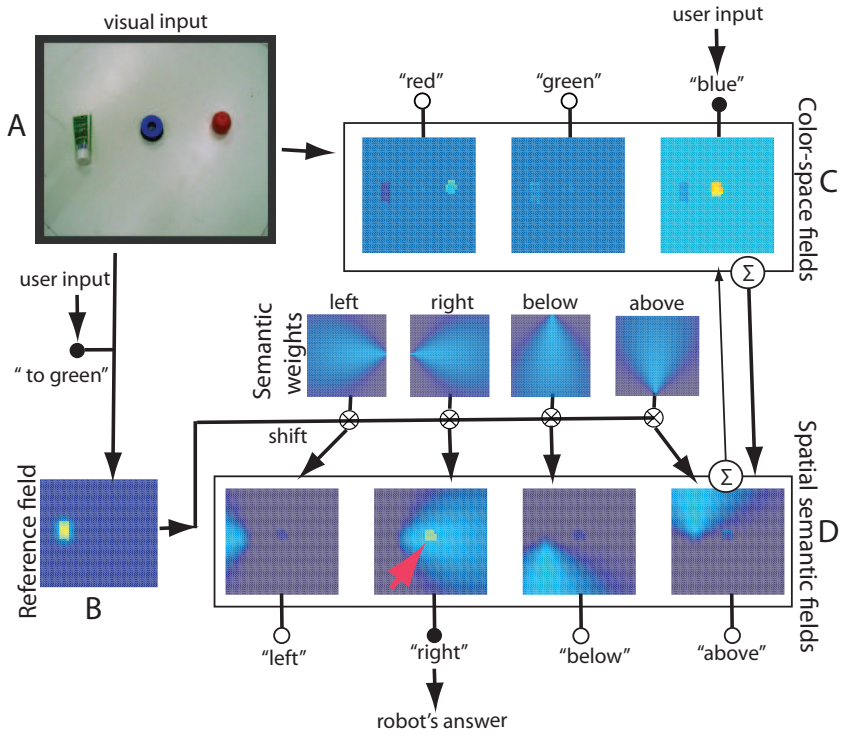


Fig. 4 Demonstration 1a. Neural fields activity just before response generation. The robot answers the question "Where is the blue object relative to the green one?" by selecting "right". **Panel A** shows the camera image (green toothpaste tube, a blue wire roll, and a red plastic apple). **Panel B** shows the reference field activation corresponding to the green reference object selected by user. **Panel C** depicts the color-space field activations induced by the current scene. The "blue" color-term node input specifying the target object uniformly raises the activation of the entire "blue" color-space field, leading to an activation peak at the blue object location. **Panel D** shows the spatial semantic field activation profiles after the shift of semantic templates to the reference object location. The active regions in the color-space fields propagate activity to the spatial semantic fields. This leads to a localized positive activation in the "right" field (red arrow) at the location of the blue target. This increases activity of the linked node, triggering the robot's answer "right".

3.1.1 Demonstration 1a: The "Where" pathway

In this demonstration, we ask "Where is the blue object relative to the green one?", by selecting the color blue for the target object and green for the reference in the user interface. Fig. 4A shows the presented visual array. The robot should select the spatial term "right". The plots in Fig. 4 show the neural fields activations just before this response.

The task input first activates the color-term node "blue". The activation of the "blue" color-term node raises the resting level of the "blue" color-space field. This uniform activation boost coupled with the camera input from the blue object induces an activation peak in the field at the location of the blue object (see "blue" Color-space field Fig. 4C). Next, the task input activates the "green" reference color-term node. This causes the green camera input to enter the reference field and induces an activation peak in the reference field representing the green item location (see Reference field, Fig. 4B). Given our emphasis on behavioral flexibility, we reiterate that there are no restrictions on the serial ordering of reference and target object color information in this sce-

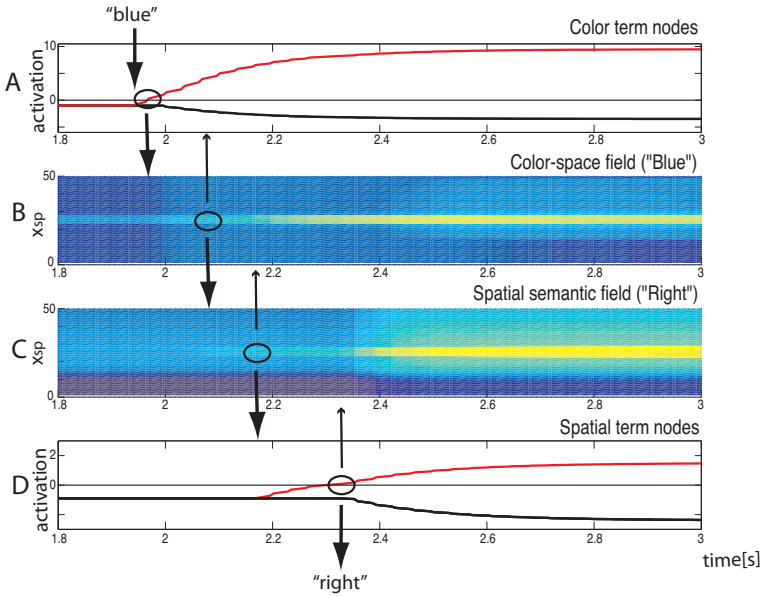


Fig. 5 Demonstration 1a time course. **Panel A** shows the color-term nodes activity over the trial (the horizontal axis represents time, the vertical axis represents activation). The "blue" input indicates the time point of linguistic input into the node. This increases activity of the "blue" color-term node (red line), causing a detection instability (ellipse) and activity propagation from the "blue" node to the "blue" color-space field (downward arrow). **Panel B** depicts the time course of the projection of the "blue" color-space field activity onto the horizontal axis over the trial. Along the vertical axis of Panel B, the lower portion corresponds to the leftmost image region, upper portion the rightmost image region. When activity in the "blue" color-space field reaches the detection instability (ellipse), that field passes activation into the spatial semantic fields (downward arrow). **Panel C** depicts the time course of the "right" spatial semantic field with activity projected onto the horizontal axis in the same manner described for Panel B. Color-space field activity leads to a localized activation profile for the blue object location (middle portion of field). Once activity surpasses the detection instability (ellipse) it propagates activation to the linked "right" spatial-term node. **Panel D** depicts the activation profile for the spatial-term nodes. The "right" spatial semantic field activity boosts the activity of the "right" node (red line), pushing it through the detection instability (ellipse), triggering the response. Smaller arrows indicate activity flow in the direction opposite to that of the dominant flow of the task. We measured time in seconds to maintain consistency across all plots in the present work. Here,$1s \approx 4 \cdot 10^3$ integration time-steps. x_{sp} is the horizontal axis of the image plane.

nario nor are there any constraints on the timing interval between these linguistic task inputs: our framework is completely flexible in this regard (see also Demonstrations 3a and 3b for probes of linguistic sequencing).

Once the target activation peak is established, the localized target activity is then transfered to the four spatial semantic fields (Fig. 4D). In addition to this vision-based input, the spatial semantic fields also receive input from the spatial semantic templates. Critically, these spatial patterns are shifted to align with the position of the reference object. Consequently, the target location activation overlaps within the "right" spatial semantic field with the semantic template (see large arrow in the "right" Spatial semantic field, Fig. 4D). This overlap ultimately leads to the activation of the associated "right" spatial-term node and thus the selection of the correct answer, "right", in the user interface.

Fig. 5 makes the time course of this task in the relevant dynamic fields more transparent. Fig. 5A presents the time course of the color-term node activation. The ellipse denotes the time of the detection instability after which

the node activity is propagated to the "blue" color-space field (approx. time=1.9s; see downward arrow). Fig. 5B shows the time course of the "blue" color-space field's activation projected onto the horizontal axes of the image plane. When the field receives the uniform activation boost from the active "blue" node (approx. time=1.9s), the activation in the field passes through a detection instability (ellipse) and begins passing input into the spatial semantic fields (see downward arrow). Within the "right" semantic field (see Fig. 5C), this input combines with the "right" spatial semantic profile which pushes activity through the detection-instability (see ellipse, Fig. 5C). Consequently, the "right" spatial semantic field then increases the activation of the "right" spatial-term node (red line, Fig. 5D), eventually moving it through the detection instability and triggering generation of the term "right" in the user interface.

3.1.2 Demonstration 1b: The "What" pathway

In this demonstration, we ask "What is the color of the object to the right of the blue one?" by selecting the spatial term "right" and the "blue" reference object color in the user interface. Fig. 6A shows the presented visual array. The robot should select the color term "red". The plots in Fig. 6 show the activation profiles just before the response.

The task input first activates the spatial-term node "right" and then the reference object color "blue". The reference object specification "blue" causes the blue camera input to enter into the reference field and induces an activation peak at the blue item location(Fig. 6B).

The activation of the "right" spatial-term node raises the resting level of the "right" spatial semantic field. This homogeneous boost creates a positive activation in this field to the right of the blue reference object once the reference information is given (see "right" Spatial semantic field, Fig. 6C). This spatially-specific activation is then input into all color-space fields. This raises activation at all those color-space field locations that lie to the right of the reference object (see lighter blue regions, Fig. 6D). Critically, this spatially-specific activation boost overlaps with the localized input from the red object in the visible scene. This overlap leads to the development of an activation peak in the "red" color-space field (see large arrow in the "red" color-space field, Fig. 6D). This stabilized peak subsequently activates the associated color-term node, triggering the correct description of the target object, "red".

Fig. 7 details the time course of this task. In Fig. 7D the time course of the "right" spatial-term node (red line) shows increased activation from the user input and the subsequent movement through the detection instability (ellipse, Fig. 7D). At this point the node begins to pass activation to the "right" spatial semantic field (upward arrow, Fig. 7D) thereby uniformly boosting the entire field. This spatial semantic field then passes through the detection-instability bifurcation (ellipse, Fig. 7C) and begins to pass activation to the color-space fields (see upward arrow into "red" color-space field, Fig. 7 B). The spatially-specific activation coming into the "red" color-space field then sums up with the localized red object activation to produce a positive activation. The field's activity thus moves through the detection instability (ellipse, Fig. 7B) to drive the activation and ultimate selection of the "red" color-term node (approx. time= 2.6s, Fig. 7A).

3.2 Demonstrations 2a and 2b: Prototypical and non-prototypical spatial relations

Demonstrations 1a and 1b illustrated the basic model behaviors, selecting either a descriptive spatial term or a color term according to the combined visual and linguistic input. This is the first demonstration of behavioral flexibilty within a single, neurally-grounded spatial language model. In both cases, however, the target object locations corresponded to perfect examples of the selected spatial terms. Empirical spatial language research, however, in-

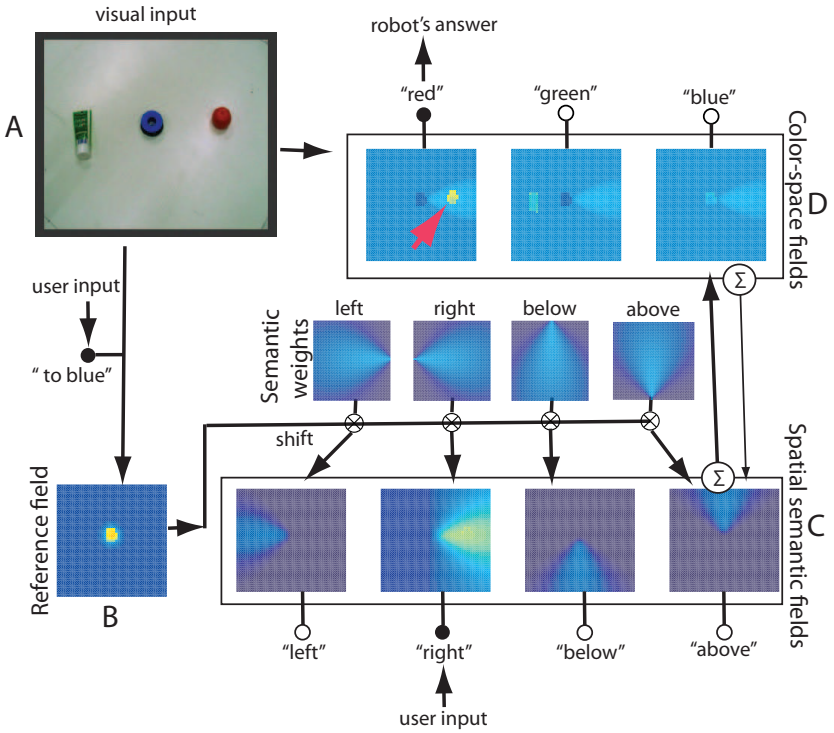


Fig. 6 Demonstration 1b neural fields activity just before response generation. The robot answers "What is the object is to the right of the blue one?" by selecting "red". **Panel A** shows the camera image (green toothpaste tube, blue wire roll, and red plastic apple). **Panel B** shows the reference field activation for the blue reference object. **Panel C** shows the spatial semantic field activation following the semantic shift to the reference object location. The "right" linguistic input boosts the entire "right" spatial semantic field. This leads to positive activation that propagates into those color-space field regions to the right of the reference object (lighter blue regions, **Panel D**). This region overlaps with that of the red plastic apple in the "red" color-space field, leading to a localized activation peak (Panel D, red arrow) which triggers the "red" response.

dicates that deviation from such prototypical spatial relations can influences spatial language decision processes (e.g. [45,15,16]). Demonstrations 2a and 2b explore the dynamic consequences of deviating from these prototypical semantic regions.

In both Demonstrations 2a and 2b, we select "blue" as the target object color and "green" as the reference object color. The robot's task in both instances then is to answer the question "Where is the blue object relative to the green one?". However, in Demonstration 2a, the relative target-reference position corresponds to a prototypical "right" relation (see Fig. 8A). In Demonstration 2b, on the other hand, the relation is neither perfectly "right" nor perfectly "above" (see Fig. 9A).

Fig. 8 shows the Demonstration 2a activities in the color-space fields (C),the reference field (B), and the spatial semantic fields (D) just before the answer is given. The spatially localized input from the robotic camera and the homogeneous boost from the blue color-term node sum to produce a localized activation peak in the "blue" color-space field (see Fig. 8C). This localized activation is then transferred to the spatial semantic fields. Here, it overlaps with the "right" spatial term template which is aligned with the reference object location (see Fig. 8D). The positive activation in the "right" spatial semantic field triggers

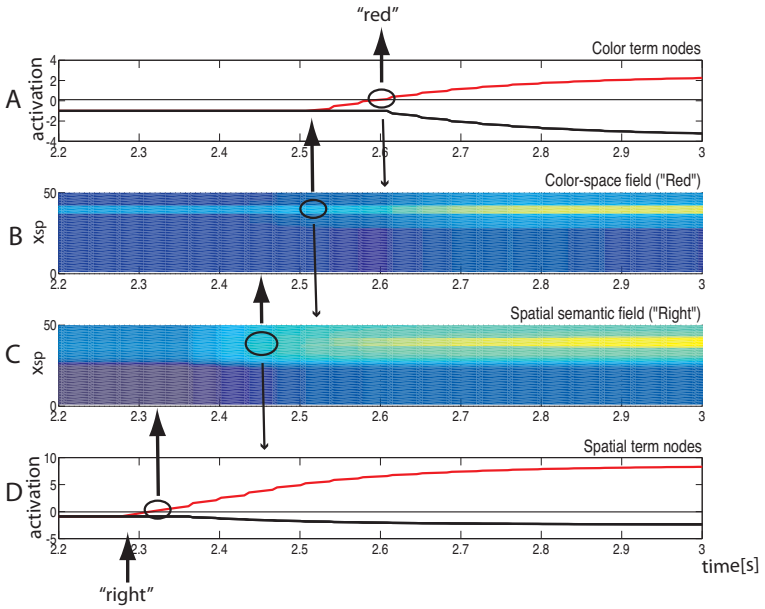


Fig. 7 Demonstration 1b time course. **Panel D** shows the spatial-term node activation over the trial (horizontal axis, seconds; vertical axis, activation). The "right" input indicates the time point of linguistic input at the start of the trial. The "right" node (red line) passes through a detection instability (ellipse), boosting the "right" semantic field (upward arrows). **Panel C** shows the "right" spatial semantic field time course (projected onto the horizontal axis as in Fig.5). Activity is elevated in the region to the right of the reference object (upper region), leading to the detection instability (ellipse) and activation into the color-space fields (upward arrow). **Panel B** shows the activation time course of the "red" color-space field (projected onto the horizontal axis). The localized activation is elevated, leading to a detection instability (ellipse). **Panel A** shows the color-term nodes activity, with the "red" color-term node (red line) triggered by the "red" color-space field activation. Smaller arrows indicate activity flow in the direction opposite to that of the dominant flow of the task.

the activation of the "right" spatial-term node, consistent with the relation in Fig. 8A.

In Demonstration 2b, we provide the same linguistic input as Demonstration 2a, but this time shift the blue target object into the upper region of the image (see Fig. 9A). As a result, the target object's spatial relation to the green referent might be best described by a combination "right" and "above" (for related empirical results see [29,45]). This semantic ambiguity is captured by the two regions of positive activation in the spatial semantic fields, one in the "right" field and the other in the "above" field (Fig. 9D). Eventually, however, the "above" field wins the competition, thereby leading to the selection of the "above" response.

These results detail how the shift of the target object's position not only changes the spatial term selected but also shapes the time course of the decision processes. As Fig. 10 shows, the response latency between the specification of the target object color and the selection of the spatial term is substantially larger in Demonstration 2b (Panels C-D, Fig. 10) than in Demonstration 2a (Panels A-B, Fig. 10). This outcome is consistent with empirical findings (e.g. [15]) showing that deviations from prototypical spatial relations can slow spatial language decision processes. By describing the competitive neural dynamics that can qualitatively capture these effects, our model provides promising grounds for addressing competitive spatial language processes and spatial term selection across varied relations.

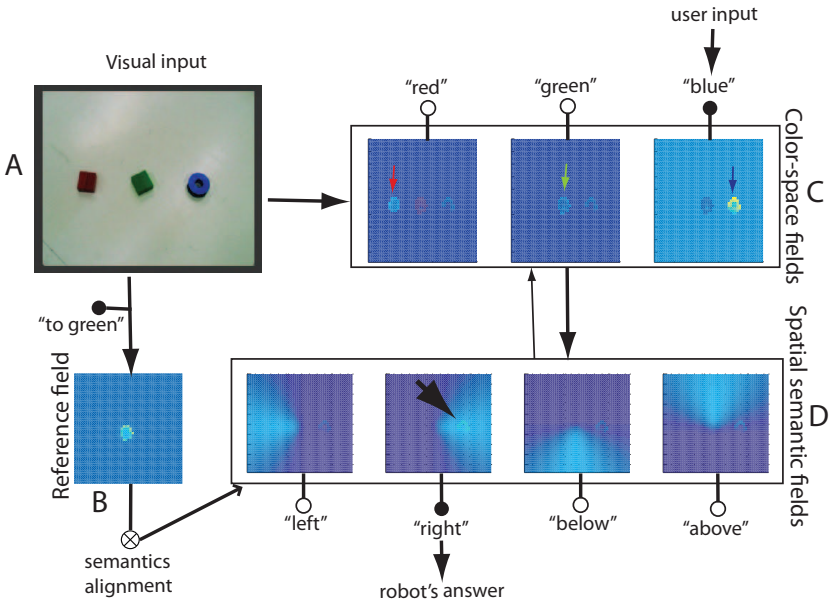


Fig. 8 Demonstration 2a neural field activity just before response generation. The robot answers "Where is the blue object relative to the green one?" by selecting "right". **Panel A** shows the camera image (a red stack, a green stack, and blue wire roll). **Panel B** shows the reference field activation corresponding to the selected green reference object location. **Panel C** shows the color-space field activity, with the "blue" color-space field boosted by the "blue" linguistic input specifying the target object. This creates a localized activation profile at the blue object location. **Panel D** shows the spatial semantic field activations which are aligned with the green reference object location and receive input from the active color-space field regions. Activation is highest in the "right" spatial semantic field which overlaps with the blue target object location (see big arrow, Panel D). This overlap leads to the activation of the "right" spatial node.

3.3 Demonstrations 3a and 3b: Dynamic signatures of linguistic sequencing

Demonstrations 1 and 2 support the sufficiency of our neural spatial language framework, revealing its capacity for behavioral flexibility and representational integration in the context of real-world visual input. Nonetheless, it should be pointed out that because the objects were differently colored, they did not share any common representational features in our model. In more complex visual environments, however, visible objects often have many features in common. As a result, unambiguously specifying an object in these environments will often require the combination of multiple descriptive terms. For example, if one is trying to specify a given red object and there are many other red objects in the scene, "The red one to the right of the blue" may suffice whereas "The red one" will clearly not.

Importantly, spoken language unfolds over time. Given that language is continuously processed [59, 1], this suggests that the sequence of words specifying an object whose features overlap with those of other objects in the scene will influence the dynamics of visual-linguistic integration. Demonstrations 3a and 3b explore the integrative dynamics of sequential linguistic input in a more complex visual environment where target identification requires both the target color and the spatial relation.

In these demonstrations, we present four items: a red stack of blocks, a green tube, a blue wire roll, and a red plastic apple (see Fig. 11A). The target object is the red stack on the left side. The robot's task is to identify the object by building an activation peak

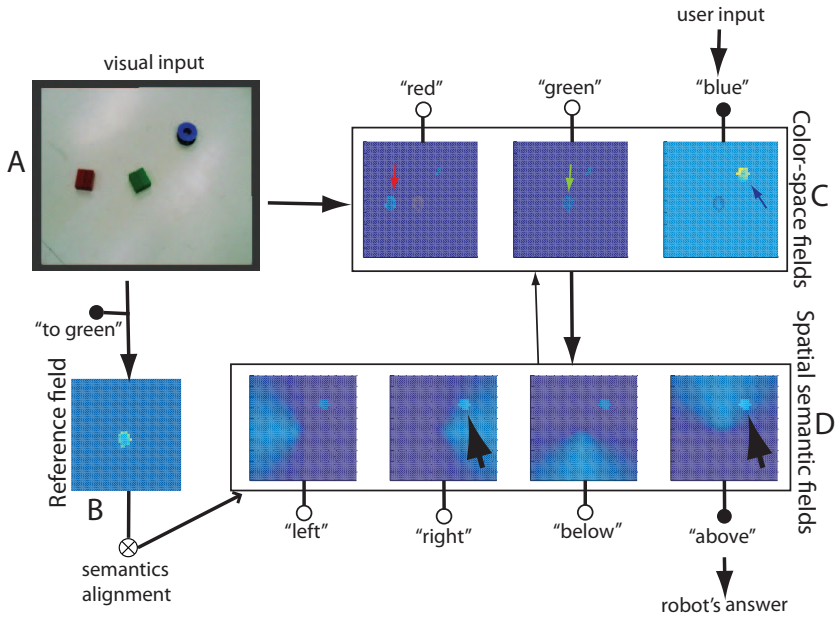


Fig. 9 Demonstration 2b neural field activity just before response generation. The robot answers "Where is the blue object relative to the green one?" by selecting the spatial term "above". The objects and the linguistic input are the same as that of Demonstration 2a but the blue target object location is shifted upwards in the image. **Panel A** shows the camera image (red stack, green stack, and blue wire roll). **Panel B** shows the reference field activation for green reference object. **Panel C** shows the color-space field activity, with the "blue" color-space field boosted by the "blue" linguistic input. This creates a localized activation profile at the blue object location. **Panel D** shows the spatial semantic field activations which are aligned with the green reference object location and receive input from the active color-space field regions. Unlike Demonstration 2a, the target location activation overlaps with both the "right" and the "above" spatial semantic fields (see large arrows). The slightly stronger overlap for the above region provides a competitive advantage eventually triggering the "above" response.

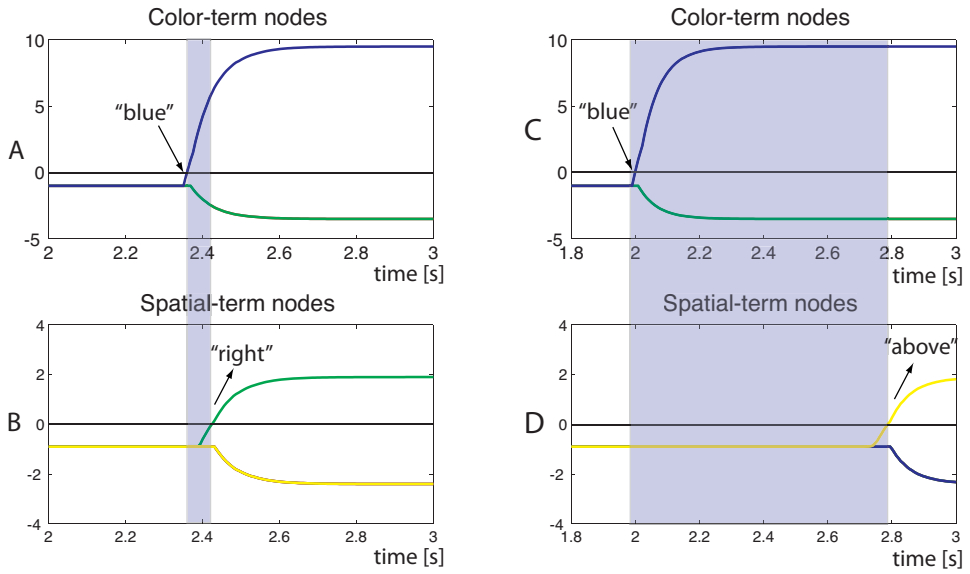
at the specified target location in the correct color-space field.

To unambiguously specify the red stack relative to the blue roll, one must give both the target object color (red) and the spatial relation (left). In line with natural speech, however, we vary the sequences. Specifically, in Demonstration 3a, we specify the spatial term ("left") first followed by the color term ("red"). In Demonstration 3b, on the other hand, the color term ("red") comes first followed by the spatial term ("left"). Although the complete descriptions are logically equivalent, differing sequences within an integrative neural dynamic model will lead to differing intermediate dynamic states. To focus on the dynamic consequences of these differing sequences, the reference object information was provided beforehand and this step is not shown.

3.3.1 Demonstration 3a: "Left" followed by "Red"

In Demonstration 3a, we first present the spatial term (left) followed by the target object color (red). This sequence roughly corresponds to describing the target as "The one to the left of the blue, the red one". The robot must build a peak at the correct target location in the correct color-space field.

As shown in Fig. 11 (left column) specifying "left" first leads to positive activation in the "left" spatial semantic field (see Fig. 11E) which is then transmitted to the color-space fields (Fig. 11D). Thus, all color-space field



(a) Demonstration 2a

(b) Demonstration 2b

Fig. 10 Demonstrations 2a and 2b time courses for the spatial-term and color-term nodes. In both demonstrations the robot answers "Where is the blue object relative to the green one?". In all panels, the vertical axis represents the node activation value and the horizontal axis represents time. **Panels A-B** show the activation for the color-term (A) and the spatial-term (B) nodes in Demonstration 2a where the target object is aligned with prototypical "right". The "blue" arrow in Panel A marks the user input specifying the target object color; the node remains active thereafter (blue line) and suppresses the other nodes. The gray region indicates the response latency between the "blue" linguistic input and the robot's selection of "right" (green line surpassing zero threshold). **Panels C-D** show the activation profile for the color term (C) and the spatial term (D) nodes in Demonstration 2b in which the target object overlaps with both the "right" and "above" regions. The wider gray bar (compare Panels A-B) indicates the greater response latency from the greater competition between the "right" and "above" spatial semantic fields.

sites to the left of the reference object receive additional input. This input overlaps with the localized visual stimuli in the "red" and "green" color-space fields because those objects both fall to the left of the referent. This gives rise to a competition between the two objects (see competing objects, Fig. 11D). At this point, the system dynamics are unstable and are largely driven by the visual input and its interaction with the spatial semantics. Because the green object is larger in the image, it maintains a slight competitive advantage over the red stack. Consequently, a peak is eventually built in the "green" color-space field at the green object location (see incorrect activation, Fig. 11D). This activation peak in turn inhibits the competing color-space fields.

When we next specify the target object color "red", however, the color-term node raises the resting level of the "red" color-space field (see Fig. 11F). The linguistic input therefore works to counteract the peak-driven inhibition from the "green" color-space field. This activation boost together with the summed, overlapping activations from the red object input and the "left" field region (Fig. 11G) leads to an activation peak at the location of the described red object(see Fig. 11F). The robot has thus selected the correct object.

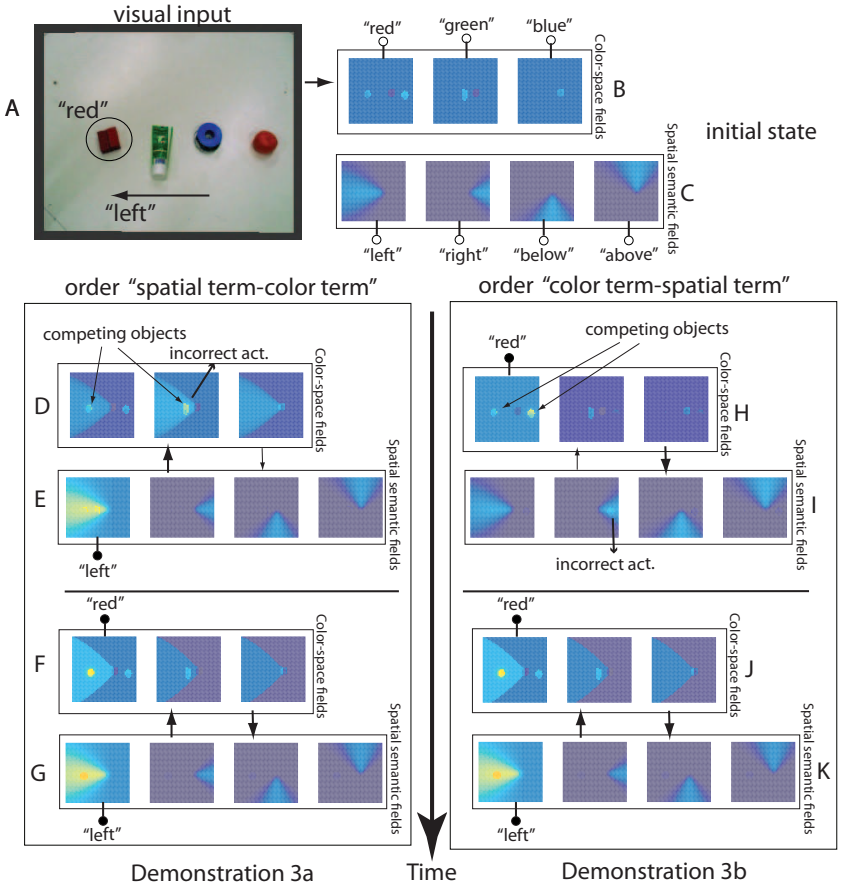


Fig. 11 Demonstrations 3a and 3b neural field activity. The robot's task is to locate the red object to the left of the blue object by building a peak at the correct location in the correct color-space field. **Panel A** shows the camera image (a red stack, a green toothpaste tube, a blue wire roll, and a red plastic apple). **Panel B** shows the initial color space field states before the target and the spatial term input are given. **Panel C** shows the spatial semantic field after semantic alignment with blue reference object. **Panels D-G (Demo. 3a, "left" then "red")**: In Panel E, the user provides the "left" linguistic input and the "left" spatial semantic field becomes more active. This activation passes to the color-space fields (Panel D) and activates the regions to the left of the reference object. This region overlaps with the green toothpaste tube and the red stack (see competing objects, Panel D). In Panel F the user provides the "red" color term input, increasing the activation of the red color-space field and creating a peak for the red stack on the left (yellow blob). **Panels H-K (Demo. 3b, "red" then "left")**: Panel H shows the increased activation from the "red" linguistic input, leading to competition between the two red objects (see competing objects, Panel H). In Panel K, the "left" linguistic input increases activation in the "left" spatial term field, boosting activity for those color-space field regions to the left of the blue reference object (Panel J) and leading to the selection of the red stack (see yellow blob in the "red" color-space field, Panel J).

3.3.2 Demonstration 3b: "Red" followed by "Left"

Demonstration 3b (Fig. 11, right column) produces a different dynamic structure. In this case we first present the "red" target color followed by the spatial term "left". This roughly corresponds the description "The red one to the left of the blue." Again, the robot must build a peak at the correct target location in the correct color-space field.

This sequence first raises the resting level of the "red" color-space field and brings the two red object locations in the field to the detection threshold (Fig. 11H). Because of the inhibitory interactions within the field, however, only one peak can be sustained. As in Demonstration 3a, the metric characteristics of the visual input drive the process. In this case, the mild shading of the red stack (see Fig. 11A) results in comparatively stronger input from the the red plastic apple. This leads to the establishment of a peak at that location which subsequently propagates activation into the corresponding "right" spatial semantic field (see incorrect activation, Fig. 11I). This in turn drives the activation of the linked "right" spatial-term node.

When we subsequently provide user input to the "left" spatial-term node (see "left", Fig. 11K), however, this activation overcomes the inhibition from the previously activated "right" node. This leads to a bifurcation and, accordingly, the "right" spatial-term node then becomes inhibited, the activity level in the "right" spatial semantic field is lowered, and that of the "left" field is raised (Fig. 11K). The elevated "left" semantic field activation in turn activates the left regions of the color-space fields (Fig. 11J), most notably in the "red" color-space field where it's activation overlaps with that from the red stack. This overlapping activation in turn creates a peak at that red stack location. The new peak therefore corresponds to the fully described target location (yellow blob, Fig. 11J).

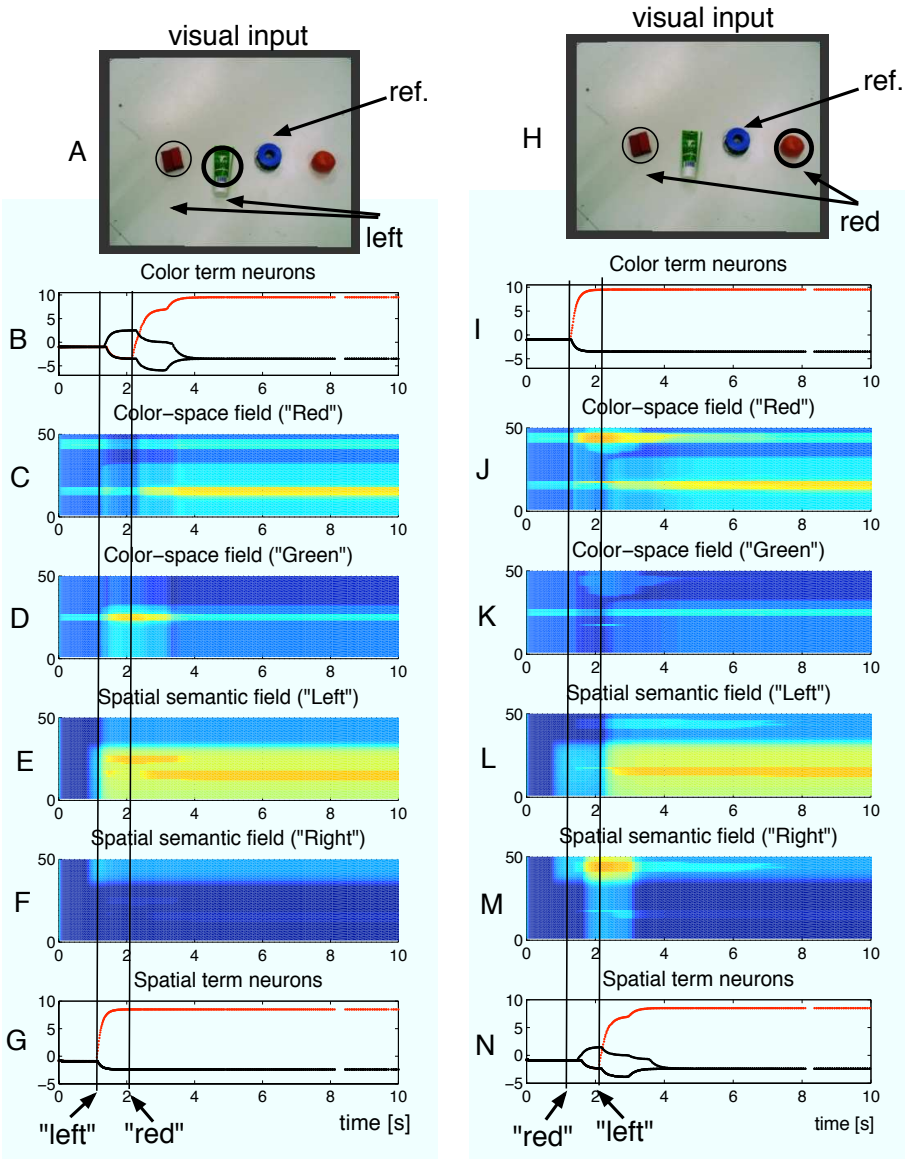
3.3.3 Linguistic Sequences: Comparing the Neural-Dynamic Time Courses

Fig. 12 further details the sequence-dependent dynamics of these tasks. Fig. 12 (left side) shows the Demonstration 3a time course which first specifies the spatial term (see arrow, Fig. 12G). This input supports the development of an activation peak in the "green" color-space field (yellow region, Fig. 12D) because it is the larger of the objects to the left of the blue referent. This activation peak in turn drives the early activation of the "green" color-term node (black line, Fig. 12B).

When we complete the description by introducing the "red" linguistic input (see time mark, Fig. 12G), however, the "red" color-term node becomes active, triggering an instability in the system and ultimately suppressing green node (Fig. 12B). This in turn facilitates the development of a localized activation peak in the "red" color-space field (yellow region, Fig. 12C). The new peak subsequently extinguishes the incorrect peak in the green color-space field (see transition from yellow to blue in Fig. 12D). The peak location shift in the "left" spatial semantic field reflects this change in the dynamic state of the system (Fig. 12E).

The Demonstration 3b time course differs dramatically (see Fig. 12, right side). The initial "red" color-term node boost increases the node activation and leads to the development of a peak at the location of the larger red object, in this case the apple to the right of the blue referent (see orange-yellow region, Fig. 12J). This subsequently builds positive activation in the "right" spatial semantic field (Fig. 12M) and activates the "right" spatial-term node (black line, Fig. 12N).

To complete the description, we then specify the "left" spatial term which triggers an instability and ultimately inhibits the previously active "right" spatial-term node (see Fig. 12N). Its positive activation also boosts the "left" spatial semantic field and enhances the activation at the red stack location in that field (note activation transitions in the "left" and "right" spatial semantic fields, Fig. 12L and M). When this "left" spatial semantic field activation propagates to the color-space fields, it increases the activation at the correct red object location and a peak emerges there (Fig. 12J). Although this is the same peak location as that in Demonstration 3a, our fine-grained analysis reveals our system's dynamic sensitivity to changes in linguistic sequencing.



(a) Demonstration 3a (b) Demonstration 3b

Fig. 12 Demonstrations 3a and 3b time courses. The horizontal axis in all panels represents time. The vertical axis in Panels B, G, I, and N represents activation level. Panels C-F and J-M project activation onto the horizontal axis; the lower region corresponds to the leftmost portion of the image, the upper region to the rightmost portion. **Panels A-G (Demo. 3a, "left" then "red")**: Panel A shows the camera image (blue wire roll reference object). In Panel G the "left" linguistic input activates the "left" node (red line), increasing activation for both objects to the left of the blue reference object. This increases activation in the "left" spatial semantic field at the green toothpaste location (see initial orange ridge, Panel E) and creates an activation peak in the "green" color-space field (Panel D); the "green" color-term node also becomes active (black line, Panel B). When the "red" linguistic input is given (red line, Panel B), however, a peak forms in the "red" color-space field (see emerging yellow activation ridge, Panel C), eliminating the "green" color-space field peak (Panel D) and shifting the activation from the green to the red object in the "left" spatial semantic field (Panel E). **Panels H-N (Demo. 3b, "red" then "left")**: The "red" color term input activates the node (red line, Panel I) and leads to an activation peak in the "red" color-space field at the red plastic apple location (first orange activation ridge, Panel J); the "right" spatial semantic field (Panel M) and "right" node also become active (black line, Panel N). When the "left" spatial term input is given (Panel N), the "left" node becomes active (red line, Panel N), increasing the "left" spatial semantic field activity in the region of the leftmost red object (orange ridge, Panel L). The increased spatial semantic activation also increases activation in the color-space field regions to the left of the reference objects. This eliminates the first activation peak in the Panel J and creates a new peak at the red stack in the leftmost portion of the color-space field (see emerging yellow-orange activation ridge, Panel J).

Notably, these integrative effects are also broadly consistent with empirical research. Spivey and colleagues [83], for example, found that people can use early linguistic information about a target object in a conjunction search task to dynamically constrain visual search processes. Eye-tracking results from Chambers and colleagues [17] reveal similar findings, showing that the presentation of constraining words like "inside" in the context of a visual scene immediately increases visual attention to those objects affording containment. Our time course analyses of linguistic sequencing differences are in line with these effects.

3.4 Demonstrations 4-5: Challenges of sensor and object movement

The previous demonstrations highlight our architecture's flexibility and robustness in the face of varying scenes and linguistic input. Movement presents an additional set of behavioral challenges. First, movements (gaze, orienting, reaching, etc) are driven by internal neural dynamic states. Thus, providing a dynamic account of emergent cognitive functions and linking these internal decision dynamics to bodily movement is an important benchmark for a viable framework.

Second, when that movement involves the sensor providing the spatial information (e.g. eyes) then the spatial relations between that sensor and the objects in the world change. Such changes in visual input and can disrupt the dynamics supporting the peaks driving cognitive behaviors. This is particularly so for spatial language where decisions depend fundamentally on spatial relations. Robustly adaptive behavior in the context of such movement is thus an important benchmark for a dynamic model of spatial language.

Finally, in addition to sensor movements, embodied cognitive systems often encounter objects moving in the world as well. Moving objects can also threaten dynamic stability because they too shift the sensory inputs supporting the peaks that drive cognitive behaviors. Generating appropriately adaptive behaviors in the context of object movements is therefore a third important test of our framework.

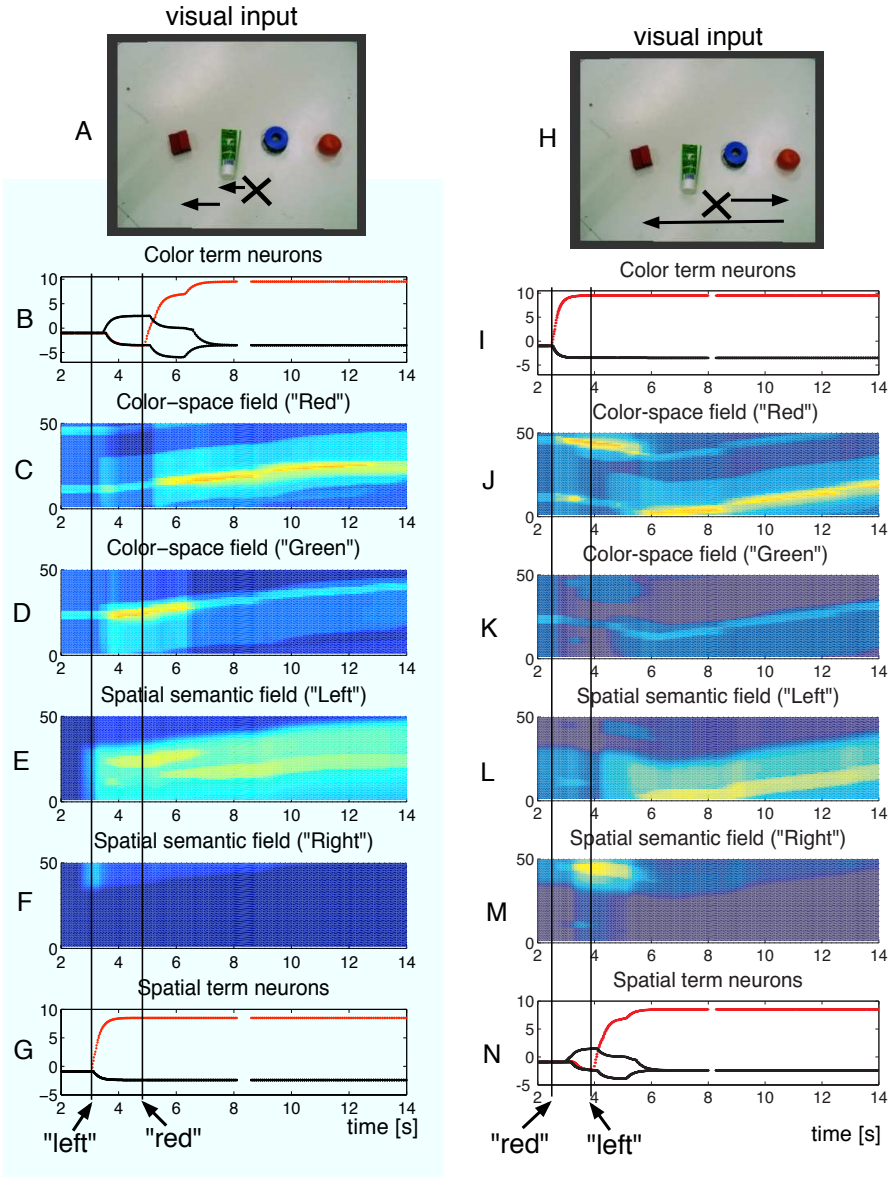
Demonstrations 4a and 4b address these first two challenges through an internally driven sensor (camera) movement. Demonstration 5 probes object movement.

3.4.1 Demonstrations 4a and 4b: Dynamically driven sensor movement

Previously discussed empirical work from Chambers and colleagues [17] indicates that eye-movements reflect the continuous integration of visual and linguistic input. To provide a behaviorally meaningful test of movement in line with the functional spirit of these results, we again probed linguistic sequencing using the same visual and linguistic input as in Demonstrations 3a ("The one to the left of the blue, the red one") and 3b ("The red one to the left of the blue").

As before, the robot's task is to build an activation peak at the specified target location in the correct color-space field. In the current movement tasks, however, we also integrated a dynamic motor control module. This module drives the robotic pan/tilt unit (see Appendix) based on the location of a peak in the color-space field, centering the corresponding object in the camera image. Movements of the camera in this context are roughly analogous to gaze shifts driven by internal dynamic processes.

Fig. 13 presents the time courses of these differently sequenced tasks (with the blue reference object already specified previously in both instances) along with the summary camera movements (see Fig. 13A and Fig. 13H). In Demonstration 4a (Fig. 13, left side) the "left" linguistic input is presented first. As we previously detailed, the green object is the larger of the two objects to the left of the reference object. This leads to a peak at that location in the "green" color-space field (see yellow region, Fig. 13D). Once this peak is established, however, the camera begins to center that location in the image by shifting to the left. This



(a) Demonstration 4a (b) Demonstration 4b

Fig. 13 Time courses of Demonstrations 4a and 4b involving camera movement. The horizontal axis in all panels represents time. The vertical axis in Panels B, G, I, and N represents activation. Panels C-F and J-M project activation onto the horizontal axis; lower region corresponds to leftmost portion of the image, upper region to the rightmost portion. **Panels A-G (Demo. 3a, "left" then "red")**: Panel A shows the camera image. The arrow next to the "X" indicates camera movement to the initially selected object; the other arrow indicates the correct item selected and centered in the image. Panel G shows initial "left" input activating the "left" spatial-term node (red line). This increases activation for both objects to the left of the blue referent. Activation at the green toothpaste location in the "left" spatial semantic field increases (see initial orange ridge, Panel E) and creates a peak in the "green" color-space field (Panel D); the "green" color-term node also becomes active (black line, Panel B). The green tube is close to the center so the camera movement is small (small shift in bounded region, Panel D). When the "red" linguistic input is given (red line, Panel B) the "red" color-space field peak forms (see emerging yellow ridge, Panel C), eliminating the "green" peak (Panel D) and shifting activation from the green to the red object in "left" semantic field (Panel E). The new peak at the described red object location drives the camera to center the selected object (see esp. Panel C). **Panels H-N (Demo. 3b, "red" then "left")**: The "red" color term input activates the node (red line, Panel I), creating a peak in the "red" color-space field at the red plastic apple (first orange ridge, Panel J); the "right" spatial semantic field (Panel M) and "right" node (black line, Panel N) also become active. This initiates a leftward camera movement (see esp. bounded region, Panel J). When the "left" input is given (Panel N), the "left" node becomes active (red line, Panel N), increasing the "left" semantic field activity by the leftmost red object (orange ridge, Panel L). This increases activation in the color-space field regions left of the referent. In Panel J, the first activation peak is eliminated and a new red stack peak emerges, driving the camera movement.

shift in turn leads to the smearing and shift of the activation profiles across all the depicted fields in Fig. 13 (left side). Nevertheless, note that this peak is stably maintained across the camera movement, thus tracking the location of the green object in the image. To this point then, our framework has shown the ability to guide the camera movement according to the specified peak location and also stably maintain that peak across the movement.

This dynamic behavioral flexibility is further born out when we then complete the description by providing the "red" color term. As discussed in Demonstration 3a, this linguistic input activates the "red" color-term node (red line, Fig. 13B) and ultimately boosts the entire "red" color-space field leading to a peak at the correct red object location (Fig. 13C). This also extinguishes the peak in the "green" color-space field (Fig. 13D). Moreover, because the specified object is even further to the left, the camera continues to shift in that direction, eventually centering the described object in the image (see centering of yellow activation profile in "red" color-space field, Fig. 13C). Again, however, this peak-driven movement does not destabilize the peak.

Demonstration 4b (Fig. 13, right side) shows comparably robust behavior for the alternative sequence in which we present the "red" color term first. In this case, the resulting uniform boost to the "red" color-space field creates an activation peak at the red apple location in the right portion of the image (see yellow activity in Fig. 13J). This in turn drives the camera to center this location in the image (see especially shifting activity profile in Fig. 13J). When we later specify the "left" spatial relation(Fig. 13N), however, this initial peak is extinguished and a peak at the fully described correct location arises instead (see later portion of Fig. 13J). This new peak then shifts the camera dynamics and the camera begins to move in the opposite direction to center the correct object (see shifting activity profiles in Fig. 13J-M).

These demonstrations together reveal our framework's ability to dynamically drive motor behaviors based on emergent neural dynamic decision processes. Moreover, they also highlight the ability to stably maintain those decisions over the resulting input shifts.

3.4.2 Demonstration 5: Target object movement

Elements in the visible world frequently move, either by their own actions (e.g. animals) or the actions of others (e.g. a person moving a coffee cup). Like sensor movements, moving objects alter the flow of visual input and therefore risk disrupting the dynamic stability on which adaptive behaviors depend. A viable neural dynamic approach to spatial language should be behaviorally robust to such movements.

To test this, we presented a blue wire-roll, a green flashlight, and a red apple (see Fig. 14A) but then moved the blue wire roll during the task. The robot's task is to identify the blue target object (blue wire roll), track its movement through the scene, and then select a descriptive spatial term when we later identify the reference object (Fig. 14B).

We began the trial by first providing input into the "blue" color-term node, thus selecting the blue wire roll as the target. We then move the blue wire roll through the visible space before specifying the reference object. Fig. 14C shows this tracking within the "blue" color-space field. After approximately six seconds, we then specify the term "green" as the reference object color, leading the robot to select the green flashlight as the referent. With the reference object specified, the spatial semantic templates become aligned with the reference location and increase the resting activation level of the relevant sites of the spatial semantic fields (see elevated activation after six seconds, Fig. 14D and E). The target object location overlaps with the activity in the "right" spatial semantic field, leading to a positive activation in this field (Fig. 14E) and trigger-

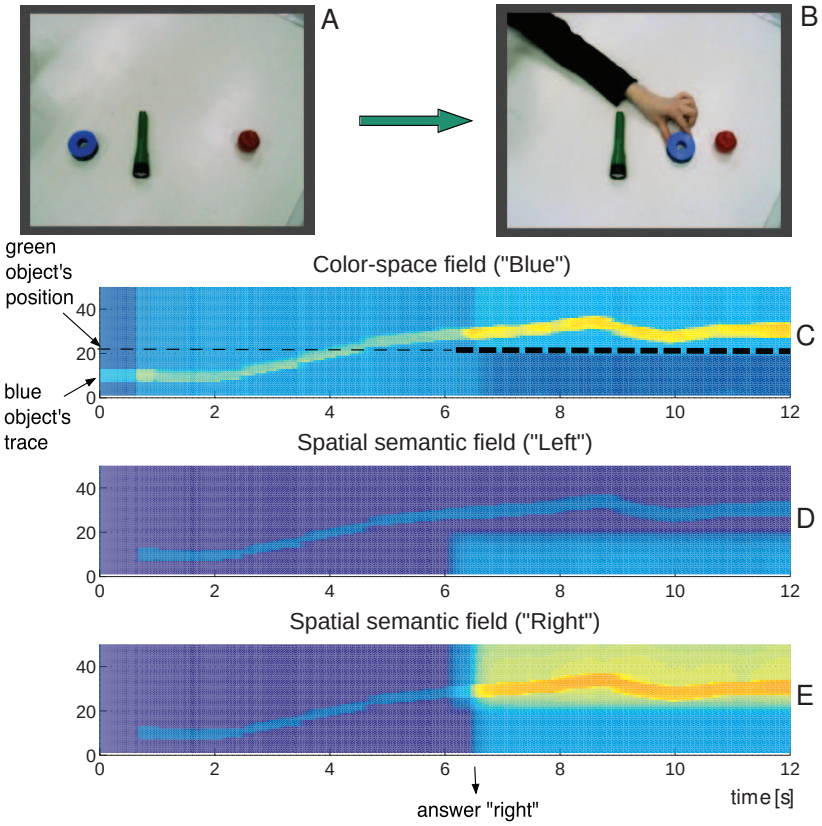


Fig. 14 Demonstration 5 time course for a moving target object. The robot's task is to track the blue target object and provide a spatial description at the final position. **Panel A** shows the workspace at the beginning of the task. **Panel B** shows the final position of the target object when the robot selects "right". In **Panels C-E**, the horizontal axis represents time; activation is projected onto the horizontal axis as in previous figures; the lower region of these fields corresponds with the leftmost portion of the image, the upper region with the rightmost portion. **Panel C** shows "blue" color space field activation. At the beginning of the trial, the blue target object is in the bottom region of the field (leftmost region of the workspace image). As the object is moved through the space, the activation profile shifts accordingly, moving eventually to the region to the right of the reference object. **Panels D and E** show the shift of the target object position through the respective "left" and "right" spatial semantic fields. When the reference object information is provided (at approximately 6 seconds), activation increases in the spatial semantic fields. The overlap between the target object location and the "right" spatial semantic region leads to an activation peak in the spatial semantic field (Panel E) and a peak in the "blue" color space field. The additional peak movements visible in the latter portions of Panels C and E arise from shaking the object and further highlight the representational robustness of the target location.

ing selection of the "right" spatial-term node. Furthermore, as an additional test of representational stability we also continued to move the target object slightly after the generation of the spatial term response (see slight variations in the peak's location in Fig. 14E after 6 seconds). The dynamic states were nonetheless stable in the face of this additional movement.

This demonstration highlights two behaviorally significant aspects of our model. First, in tracking the target item, the robot again stably represented the target object location despite the substantial change in visual input and even tolerated the potentially disruptive presence of the hand in the scene. This further extends the behavioral flexibility of the system because it provides another instance of

successful operation in an unstable, variable environment. Second, by tolerating an object movement in the period between the linguistic inputs specifying the task, the robot also again displayed the ability to successfully integrate linguistic and visual input over time. This complements the sensor movement demonstrations and further substantiates our model's ability to dynamically structure behavior in the presence of non-static visual input.

4 Discussion

4.1 Summary

Language behaviors are generated by real bodies in real time. To facilitate the development of a formalized theoretical framework for embodied language processes, we implemented a multi-component neural dynamic model emphasizing behavioral flexibility and representational integration in spatial language. Tests of our architecture implemented on a robotics platform across five different demonstration sets using real visual input support the viability of this approach.

In Demonstration 1a we first tested the "where" pathway by asking "Where is the blue object relative to the green one?" in the context of a three-item visual scene. The system autonomously selected the correct spatial term "right" using only this input. This task captures two of our three key spatial language characteristics. First, in correctly applying the color-based descriptions to the visual scene and generating a spatial term, the system necessarily integrated spatial and non-spatial (color) representations. Second, in combining the localist node activation with the color-specific visual input, our model also demonstrated the capacity to integrate symbolic functionality with continuous, graded spatial representations. The subsequent probe of the "what" pathway in Demonstration 1b ("Which object is the right of the blue?") produces a similar verification, again showing how a graded, neurally-grounded approach to scene representations can produce integrated representations across feature dimensions and between the symbolic and the continuous.

These two demonstrations also represent two qualitatively different behaviors, one generating a spatial term ("right") from descriptions of the target and the referent, the other extracting a target object color ("red") from a description of the referent and a spatial relation. This represents the first evidence of behavioral flexibility in our model. As such, it is important to emphasize that only the contextually-specific input differed between these cases. These behaviors did not require different parameter values nor did they require an external controlling input. This behavioral structuring instead inheres in the underlying autonomous processes and their continuous coupling to sensory inputs.

The behaviors in Demonstrations 1a and 1b depend fundamentally on the neural dynamic concepts of autonomy, gradedness, and stability. In building on these same neural concepts, Demonstrations 2a and 2b also made contact with metric spatial language effects. Specifically, we showed that shifting a target away from a prototypical spatial relation (e.g. right) to a non-prototypical one increased the degree of competition within the system and the response generation time. This result is generally consistent with empirical spatial language research (e.g. [15]) and is generic within the neural dynamics and thus captures a general principle.

Although these initial tests alone show that a neurally grounded approach can address core aspects of spatial language, our subsequent results show that these basic functions in fact enable a far greater behavioral breadth. Consider Demonstrations 3a and 3b, the four-item scenarios. Correctly identifying the target object in these tasks required both color and spatial term information. However, this information was provided sequentially (as it would be with natural speech), not simultaneously. Providing the symbolic information sequentially as we did in the four item scenarios thus tested the

model's ability to continuously integrate information as it becomes available yet still arrive at the correct answer.

The results suggest that our framework is indeed substantially tolerant of such timing variability. Demonstration 3a, for example, provided the spatial term first before the disambiguating color term, thus leading an early, but incorrect, target selection. When the target-specific color information was later provided, however, this new information was then incorporated and the correct target peak was established. In Demonstration 3b, on the other hand, we reversed the order by specifying the target object color first which again lead to the generation of an incorrect target peak. Despite this radically different time course, the system nevertheless again eliminated the incorrect peak and created a peak for the correct target object once provided with the disambiguating spatial term.

This is a particularly strong test of linguistic input timing tolerance for two reasons. First, across the demonstrations, we reversed the order of the color and spatial term information, not simply the timing interval within some fixed linguistic input sequence. Second, within each demonstration, the dynamic interplay between the currently available linguistic input and the slight activation advantage for the larger of the two possible targets led to an initial, incorrect answer. Nevertheless, despite the inhibition created by such peaks, the system still ultimately created the correct target object peak. This draws attention to a new dimension of behavioral flexibility, namely consistent cognitive decision processes in the face of highly variable linguistic input. Importantly, this tolerance to linguistic input variability is not the product of an "insensitive" dynamic system. That is, instead of preventing the emergence of peaks in all but the most fully specified scenarios, our system instead maintains the ability to flexibly build peaks based on partial information along with the ability to build new, competing peaks as information unfolds. As we previously noted, this accords well with empirical research demonstrating the continuous integration of the visual and linguistic inputs [17, 83, 86].

This empirical evidence of continuously integrative language processing, namely eye-tracking research, draws attention to another challenge for our system. A core premise of these and other eye-tracking studies is that motor behaviors (e.g. eye-movements) reflect underlying cognitive states. The ability to adaptively structure motor behaviors, specifically camera movements, according to the model's internal neural dynamic states is thus a significant test for our framework. In initiating such sensor movements, however, our implementation must also provide for the representational stability; in the absence of such stability the shifting spatial relations between the sensor and the objects in the world could perturb the scene representations that support adaptive, flexible behavior.

Demonstrations 4a and 4b addressed both these challenges by again presenting sequentially varying linguistic inputs but now driving camera movement from the neuronal dynamics. Results from these two scenarios showed that camera movements changed according to the internal dynamic states of the system. Moreover, in eventually shifting from the incorrect target to the fully specified correct target, we also demonstrated a robust tolerance for changes in the visual stimuli. Furthermore, in integrating spatial and non-spatial features as well as symbolic and continuous representations, our system generated another, wholly embodied behavior – movement. This capacity emerges directly from our neurally-based approach to symbol grounding and our attention to the core neural concepts of autonomy, gradedness, and stability.

In Demonstration 5, our final test, we further examined our system's robustness to movement, this time by shifting the object before providing the reference object color. As a consequence, the system needed to track the specified target object through space and time before receiving the reference object color and

selecting the correct spatial term. Nevertheless, our model again accurately integrated the sequential linguistic input and stably maintained the scene-symbol link required to answer correctly. This result further substantiates the model's capacity for representational stability and the behavioral flexibility conferred by a neurally-based approach to scene representations.

4.2 Neural foundations

The theoretical language we used is grounded in the following neural principles which play a central role in our account. (1) All representation is based on graded activation variables. (2) Space, perceptual features, and movement parameters are captured by continuous dimensions along which activation fields are defined. The principle of neural fields reflects the encoding of such information by populations of neurons, whose feedforward path from the sensory surface or to the motor surface determines how they contribute to the activation fields. (3) The neural dynamics that characterize the temporal evolution of the activation fields consist of (a) external inputs, which mediate both feedforward connectivity from sensory input and the coupling among different fields, and (b) intra-field interaction, which reflects the generic cortical pattern of local excitation and global inhibition. The stability of local activation peaks, which are the units of representation, emerges from this pattern of interaction. (4) When peak locations are not specified by inputs but by learned patterns of excitability, neural fields act like categorical neural representations described by individual dynamic neurons.

Adopting these neural principles as constraints for the theoretical modeling of spatial language is a necessary, but not yet sufficient, condition for a comprehensive neural account of spatial language behaviors. One may envisage a further step, however, in which specific populations in particular parts of the higher nervous systems are assigned particular functions. Because we know much about the early visual system, particularly its representation of retinal space and perceptual features, it is easy to envision broad qualitative assignments. Neural correlates of object perception and recognition, for example, have been found in the ventral stream [85]. In addition, an initial understanding of how parietal structures in the dorsal stream, in particular, LIP, may enable more abstract, object-centered spatial representations [97] is also emerging. The neurophysiological foundations of goal-directed reaching movements have also been extensively studied [32]. The neural mechanisms of language and speech are known at a much more macroscopic level, however [68].

While promising, we believe that this next step is still outside the range of current neurophysiological research. This belief is partly driven by practical considerations. In particular, the broad diversity of neural functions invoked in spatial language has not been studied at a consistent level of resolution across the many potentially involved brain areas. This belief is also driven in part by theoretical and conceptual considerations. Specifically, the neural dynamics that provide the requisite stability and coupling are strongly interactive. Such interaction makes assignment of neural function to particular substructures particularly difficult. When a subpopulation of the neural dynamic system is removed, for instance, a particular cognitive function may fail to emerge. Such failure need not, however, imply that the subpopulation in question "is responsible for" that particular function. The failure may instead come about because input from the removed subpopulation to another subpopulation is now missing. This missing input might then prevent that other subpopulation from reaching the dynamic regime needed to stabilize the neural representations critical for the relevant function. At the same time, the input needed may also be quite non-specific to the neural function. It could, for example, be

something as generic as a constant or a broad input that enables peak generation.

4.3 Connections with Established Spatial Language Research

Our framework adopts a qualitative neural-dynamic approach to spatial language, emphasizing the continuous integration of sensory-motor processes and linguistic input. In doing so, it aims to address a host of issues typically overlooked in spatial language theories to date, including representational integration and behavioral time courses. For this reason, we believe that our process-based approach is uniquely well placed to address flexible spatial language behavior.

While theoretically distinct, our work does nevertheless have a strong connection with the established spatial language literature. Our spatial semantics, for example, are implemented with a separate set of connection weights which are dynamically aligned with a reference object and applied to a visual scene. This is conceptually similar to the notion of spatial templates developed by Logan and colleagues [58, 16,15] in which spatial regions are divided according to good, acceptable, and bad instances of a spatial relation term. Our spatial semantic approach may therefore be described as a dynamic instantiation of this idea. Additionally, the partial overlap of our semantic fields is also consistent with empirical work showing that some spatial locations are best described with a combination of spatial terms (e.g. above and a little to the right) rather than a single, exclusive term [45,29].

Our framework also captures some core elements of the spatial apprehension sequence from Logan and colleagues [16,15,57]. These elements conceptually outline the steps individuals take to confirm the presence or absence of a described spatial relation in a visual scene (e.g. "the dash is above the plus"), including indexing the arguments of the spatial relation (the target and reference objects) onto the visual scene, establishing the reference frame within the scene, and applying the specified semantics accordingly. Although our framework does employ some simplifying assumptions, namely a single reference frame at a fixed rotation, it does show how the spatial indexing and semantic application steps may be instantiated within a neural dynamic framework. It also extends these basic steps to a broader range of tasks and therefore shows how these functions may be accounted for within a behaviorally flexible framework that can tolerate sequence variation. The core elements described by Logan and colleagues thus appear to be conceptually primary in spatial language although our results reveal that their dynamic details can vary considerably according to the specific visual and linguistic context.

4.4 Limitations

In order to focus on representational integration and flexibility in spatial language, we made some simplifications in our model. Some resulting limitations bear noting. First, as we alluded to earlier, we simplified our reference frame alignment process. We assumed a single default viewpoint and thus sidestepped the complexities of reference frame rotation. As a result, our work cannot address evidence that spatial language behaviors are sensitive to changes in the rotation of reference objects with canonical orientations (e.g. chair; [12,14,16]).

Our object representations were also simplified. This is tied to the reference frame rotation issue because alternative, object-based intrinsic reference frames require object orientation information. In acknowledging this limit, however, we also note the recent development of a dynamic field model of object recognition which can quickly learn to recognize multi-feature objects [28]. The inclusion of orientation information in these representations suggests that our approach is well suited to incorporating more complex object representations.

Parsing of the input stream is another element absent in our framework. However, re-

cent models show that neurally grounded approaches can parse linguistic streams [46] and embed symbolic parsing processes within temporally continuous neural dynamics [41]. The conceptual mapping from these dynamics onto the DFT is therefore feasible.

Finally, although stability plays a central role in the generation of behaviors across the tasks, particularly the movement scenarios, there is one case in which stability is a problem. Specifically, after generating a spatial or color term response, the linguistic node stays at the same activation level. This stability prevents the generation of a new decision. This would be a problem if, for instance, the system selected a descriptive term after which the target was moved to a new spatial relation requiring a different spatial term. In essence, the system fails to register the generation of its own response and thus cannot shift to a qualitatively different state as a result. This is inconsistent with our emphasis on the continuously adaptive structuring of behavior. One could address this with a form of contextually-dependent feedback that effectively recognizes the generation of a linguistic response. This points to the need for greater behavioral organization that generalizes beyond the spatial language scenarios. Although no comprehensive approach to behavioral organization yet exists, recent work does suggest that neural dynamic theories like the DFT might also provide the grounds for developing this capability [72].

5 Conclusions

We began with the observation that theoretical treatments of language are often dissociated from the unfolding of behavior in real-time. In order to address this problem and build on the growing empirical support for a real-time, embodied foundation of language we adopted a systems-level neural dynamic perspective. We brought these theoretical tools to bear on spatial language, a domain that directly connects linguistic processes and the sensory-motor surfaces embedded in the world. With this vantage point we further proposed that addressing the neural dynamic processes supporting scene representation could provide the basis for a behaviorally flexible spatial language system. To test this claim, we developed a neural dynamic architecture grounded in the Dynamic Field Theory and implemented it on a robotics platform linked to a real-time camera image of a shared workspace.

Our results show that attending to the neural dynamic details of scene representation can provide the foundation for flexible, contextually-dependent spatial language behaviors. Across the demonstrations our model generated differing responses based solely on the linguistic and visual input. These outcomes reveal the system's capacity for representational integration and the grounding of linguistic terms in dynamic sensory-motor processes. They also verified the ability to behave continuously and dynamically integrate new linguistic information as it unfolds. Furthermore, these demonstrations also shed light on how neural dynamic scene representations and variability in the strength of the visual input can shape the time course of these behaviors.

Our framework has important theoretical consequences both within the spatial language domain and for language research more generally. In the realm of spatial language, this work is the first demonstration of behavioral flexibility within a unified, neurally grounded theoretical framework. Thus, while it is certainly not the first robotics platform dealing with the complexities of real-world visual and linguistic input (e.g. [63]) it is to our knowledge the first to do so in a manner aligned with the neural dynamic foundations of embodied cognition. Applied to the broader domain of language more generally, this work therefore highlights the power of attending to the fine-grained dynamic details of non-linguistic processes supporting "higher" level language. Our demonstrated satisfaction of several key constraints across several complex and varied scenarios also suggests that systems-level neural

dynamic theories, such as the DFT, can provide the conceptual foundation needed to unite real-time language and the sensory-motor world.

6 Appendix

6.1 Color-space fields

Conceptually, the three-dimensional color-space field evolves according to the equation:

$$\begin{aligned}\tau\dot{U}_{col}(c,x,y,t) &= -U_{col}(c,x,y,t) + h \\ &+ I(c,x,y,t) + \int f\Big(U_{col}(c',x',y',t)\Big) \\ &\times \omega(c-c',x-x',y-y')dc'dx'dy'\end{aligned} \tag{5}$$

In the implementation, however, we resolved the color dimension sparsly, substituting the interactions in this dimension through the global inhibition between the six two-dimensional color-space fields:

$$\begin{aligned}&\tau\dot{U}_c(x,y,t) = -U_c(x,y,t) + h + I(c,x,y,t) \\ &+ \int f\Big(U_c(x',y',t)\Big)\omega(x-x',y-y')dx'dy' \\ &- \sum_{c'\neq c}\int_{x',y'} f\Big(U_{c'}(x',y',t)\Big)dx'dy', \\ &c = 1..N_{col} = 6\end{aligned} \tag{6}$$

The spatial interaction kernel was the sum of a Gaussian for the local excitation and a global inhibitory term:

$$\begin{aligned}&\omega(x-x',y-y') = \\ &c_{exc}\exp\Big(\frac{(x-x')^2+(y-y')^2}{\sigma_{exc}^2}\Big) - c_{inh}\end{aligned} \tag{7}$$

The sigmoidal non-linearity smoothing the output of the fields was:

$$f\Big(U_c(x,y,t)\Big) = \frac{1}{1+\beta|U_c(x,y,t)|}, \quad \beta = 80 \tag{8}$$

The parameters were:

$$\begin{aligned}\tau &= 10 \\ h &= -2 \\ c_{exc} &= 0.3 \\ \sigma_{exc} &= 3 \\ c_{inh} &= 1\end{aligned}$$

The external input to the color-space field was formed as:

$$\begin{aligned}&I(c,x,y,t) = c_{cam}I_{cam}(c,x,y,t) \\ &+ c_{sp}I_{sp}(x,y,t) + c_{node}I_{node}(c,t) \\ &I_{sp}(x,y,t) = \sum_{i=1}^{N_{sp}} f\Big(U_{sp}(x,y,t)\Big) \\ &I_{node}(c,t) = f\Big(d_c(t)\Big) \\ &c_{cam} = 4 \\ &c_{sp} = 2 \\ &c_{node} = 3\end{aligned} \tag{9}$$

The input from the visual sensor $I_{cam}(c,x,y,t)$ was formed as described in section 2.2.1.

The color-space fields were represented as 50×50 matrices for calculations.

6.2 Color-term nodes

The dynamical nodes' activity evolved according to equation

$$\begin{aligned}\tau\dot{d}_c(t) &= -d_c(t) + h_d + c_{exc}f\Big(d_c(t)\Big) \\ &-c_{inh}\sum_{c'\neq c} f\Big(d'_c(t)\Big) + I_d(c,t)\end{aligned} \tag{10}$$

The external input to a node was defined as:

$$\begin{aligned}&I_d(c,t) = c_{GUI}I_{GUI}(c,t) \\ &+ c_U\int_{norm} f\Big(U_c(x,y,t)\Big)dxdy\end{aligned} \tag{11}$$

Here, $\int_{norm}$ denotes the summed activity in the respective field normalized by the filed's size (divided by $x_{max}y_{max}$). $I_{GUI}(c,t) = 1$, if c=color selected by the user; $I_{GUI}(c,t) = 0$ otherwise.

The parameters were:

$$\tau = 10$$
$$h_d = -1$$
$$c_{exc} = 0.5$$
$$c_{inh} = 2.5$$
$$c_{GUI} = 7$$
$$c_U = 3$$

6.3 Reference field

The reference field evolved according to the dynamics

$$\tau \dot{U_R}(x,y,t) = -U_R(x,y,t) + h + I_{cam}(x,y,t) + \int f\Big(U_R(x',y',t)\Big)\omega(x-x',y-y')dx'dy' \tag{12}$$

The spatial interaction kernel and the sigmoidal non-linearity were the same as for the color-space fields. The parameters were:

$$h = -1$$
$$c_{inh} = 0.5$$
$$c_{exc} = 0.3$$
$$\sigma_{exc} = 3$$

The camera input $I_{cam}(x,y,c=c_{ref},t)$ was formed as described in the main text (section 2.2.3). The color c_{ref} of the refererence object was specified by the user.

6.4 Semantic templates

The semantic template functions were:

$$M_{sp} = \exp\left[-\frac{(\rho-\rho_0)^2}{2\sigma_\rho^2}\right]\exp\left[-\frac{(\theta-\theta_0)^2}{2\sigma_\theta^2}\right] \tag{13}$$

where
$\sigma_\rho = 40$ (at fields' size 50×50),
$\sigma_\theta = 60$ rad,
$\rho_0 = 5$,
$\theta_0 = \pi, 0, \pi/2, -\pi/2$
(for "left", "right", "above", "below" respectively).

To reduce the computational overload, the weight matrices were represented centered on the edge of the matrix. When convolving with the output of the reference field, the weight matrices (used as kernel) were anchored accordingly.

6.5 Spatial semantic fields

Spatial semantic fields evolved according to equation

$$\tau \dot{U}_{sp}(x,y,t) = -U_{sp}(x,y,t) + h + I_{sp}(x,y,t) + \int f\Big(U_{sp}(x',y',t)\Big)\omega(x-x',y-y')dx'dy', \quad sp = 1..N_{sp} = 4 \tag{14}$$

The spatial interaction kernel and the sigmoidal non-linearity were the same as for the Color-space fields.

The parameters were:

$$h = -5$$
$$c_{inh} = 0.05$$
$$c_{exc} = 0.3$$
$$\sigma_{exc} = 3$$

The external input

$$I_{sp}(x,y,t) = c_c \sum_{i=1}^{N_{col}} f\Big(U_c(x,y,t)\Big) + c_{shift} I_{shift,sp}(x,y,t) + c_{snode} f\Big(d_{sp}(t)\Big), \quad c_c = 2.2, \quad c_{shift} = 0.2, \quad c_{snode} = 4.5 \tag{15}$$

The $I_{shift,sp}$ was the result of the "shift" operation, aligning each of the spatial semantic templates with the location of the reference object.

6.6 "Shift"

The shift was accomplished by convolution of the outcome of the reference field with the spatial semantic templates:

$$I_{shift,sp}(x,y,t) = \int_{x',y'} f\Big(U_R(x',y',t)\Big) \times \\ \times M_{sp}(x-x', y-y', t)dx'dy' \tag{16}$$

For computational efficiency we approximated this integral by summation over the points of positive activation in the reference field.

6.6.1 Short introduction to the navigation dynamics

The camera head is mounted on a robotic pan-tilt unit, which can be controlled via a PVM (parallel virtual machine) interface directly from our software. A navigation module implemented the dynamic navigation (proposed in [10]) but without the obstacle avoidance component. The target contribution was calculated from the summed output activation in the color-space fields (17).

$$\begin{aligned} T(x,y,t) &= \sum_{i=1}^{N_{col}} f\Big(U_c(x,y,t)\Big) \\ F_{pan}(t) &= R(x) \cdot T_x(x,y,t) \\ F_{tilt}(t) &= R(y) \cdot T_y(x,y,t) \\ \tau \dot{Pan}(t) &= -Pan(t) + F_{pan}(t) \\ \tau \dot{Tilt}(t) &= -Tilt(t) + F_{tilt}(t) \end{aligned} \tag{17}$$

Here, $R(x) = x - \frac{x_{max}}{2}, R(y) = y - \frac{y_{max}}{2}$ were monotonic functions defining the mapping between the distance from the positive activation in the color-space field to the center of the fields and the strength of the attractor associated with the positive activation.

Thus, as soon as positive activation signaled the detection of the object of interest in the visual array, the head moved smoothly to center that object. Because the representation of objects in the color-space fields was updated , the color-space fields effectively tracked the visual scene. The dynamics of the whole framework was also autonomously updated.

References

1. Allopenna, P., Magnuson, J., Tanenhaus, M.: Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory & Language **38**, 419–439 (1998)
2. Amari, S.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics **27**, 77–87 (1977)
3. Andersen, R.A.: Multimodal representation of space in posterior parietal cortex and its use in planning movements. Annual Reviews of Neuroscience **20**, 303–330 (1997)
4. Bangerter, A.: Using pointing and describing to achieve joint focus of attention. Psychological Science **15**, 415–419 (2004)
5. Barsalou, L.W.: Perceptual symbol systems. Behavioral and Brain Sciences **22**, 577–660 (1999)
6. Basole, A., White, L.E., Fitzpatrick, D.: Mapping multiple features in the population response of visual cortex. Nature **423**, 986–990 (2003)
7. Bastian, A., Riehle, A., Erlhagen, W., Schöner, G.: Prior information preshapes the population representation of movement direction in motor cortex. NeuroReport **9**, 315–319 (1998)
8. Bastian, A., Schöner, G., Riehle, A.: Preshaping and continuous evolution of motor cortical representations during movement preparation. European Journal of Neuroscience **18**, 2047–2058 (2003)
9. Beer, R.: Dynamical approaches to cognitive science. TRENDS in Cognitive Sciences **4**, 91–99 (2000)
10. Bicho, E., Schöner, G.: The dynamic approach to autonomous robotics demonstrated on a low-level vehicle platform. Robotics and autonomous systems **21**, 23–35 (1997)
11. Cangelosi, A., Riga, T.: An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. Cognitive Science **30**, 673–689 (2006)
12. Carlson, L.: Selecting a reference frame. Spatial cognition and computation **1**, 365–379 (1999)
13. Carlson, L.: Attention unites form and function in spatial language. Spatial Cognition and Computation **6**, 295–308 (2006)
14. Carlson, L.: Inhibition within a reference frame during the interpretation of spatial language. Cognition **106**, 384–407 (2008)
15. Carlson, L., Logan, G.: Using spatial terms to select an object. Memory and Cognition **29**, 883–892 (2001)

16. Carlson-Radvansky, L.A., Logan, G.D.: The influence of reference frame selection on spatial template construction. Journal of Memory and Language **37**(3), 411–437 (1997)
17. Chambers, C., Tanenhaus, M., Eberhard, K., Filip, H., Carlson, G.: Circumscribing referential domains during real-time language comprehension. Journal of Memory and Language **47**, 30–49 (2002)
18. Cisek, P., Kalaska, J.F.: Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. Neuron **3**(45), 801–814 (2005)
19. Clark, H.: Using language. Cambridge University Press, Cambridge (1996)
20. Colby, C.: Action-oriented spatial reference frames in cortex. Neuron **20**, 15–24 (1998)
21. Coventry, K., Cangelosi, A., Rajapakse, R., Bacon, A., Newstead, S., Joyce, D., Richards, L.: Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In: C. Freksa (ed.) Spatial Cognition IV, vol. LNAI 3343, pp. 98–110. Springer-Verlag, Heidelberg (2005)
22. Crowe, D., Averback, B., Chafee, M.V.: Neural ensemble decoding reveals a correlate of viewer- to object-centered spatial transformation in monkey parietal cortex. The Journal of Neuroscience **28**(20), 5218–5228 (2008)
23. Damasio, H., Grabowski, T., Tranel, D., Ponto, B., Hichwa, R., Damasio, A.R.: Neural correlates of naming actions and of naming spatial relations. Neuroimage **13**, 1053–1064 (2001)
24. Denis, M., Pazzaglia, F., Cornoldi, C., Bertolo, L.: Spatial discourse and navigation: An analysis of route direction in the city of venice. Applied Cognitive Psychology **13**, 145–174 (1999)
25. Elman, J.: On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. Cognitive Science **33**, 547–582 (2009)
26. Erlhagen, W., Bastian, A., Jancke, D., Riehle, A., Schner, G.: The distribution of neuronal population activation (dpa) as a tool to study interaction and integration in cortical representations. Journal of Neuroscience Methods **94**, 53–66 (1999)
27. Erlhagen, W., Schöner, G.: Dynamic field theory of movement preparation. Psychological Review **109**, 545–572 (2002)
28. Faubel, C., Schöner, G.: Learning to recognize objects on the fly: a neurally based dynamic field approach. Neural Networks **21**, 562–576 (2008)
29. Franklin, N., Henkel, L.: Parsing surrounding space into regions. Memory and Cognition **23**, 397–407 (1995)
30. Gardner, J., Merriam, E., Movshon, J., Heeger, D.: Maps of visual space in human occipital cortex are retinotopic, not spatiotopic. The Journal of Neuroscience **28**, 3988–3999 (2008)
31. Gegenfurtner, K.: Cortical mechanisms of colour vision. Nature Reviews Neuroscience **4**, 563–572 (2003)
32. Georgopoulos, A.P.: Neural aspects of cognitive motor control. Current Opinion in Neurobiology **10**(2), 238–241 (2000)
33. Georgopoulos, A.P., Schwartz, A.B., Kettner, R.E.: Neural population coding of movement direction. Science **233**, 1416–1419 (1986)
34. Glenberg, A.: What memory is for. Behavioral and Brain Sciences **20**, 1–55 (1997)
35. Glenberg, A., Kaschak, M.: Grounding language in action. Psychonomic Bulletin and Review **9**, 558–565 (2002)
36. Glenberg, A., Robertson, D.: Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. Journal of Memory and Language **43**, 379–401 (2000)
37. Gliozzi, V., Mayor, J., Hu, J.F., Plunkett, K.: Labels as features (not names) for infant categorization: A neurocomputational approach. Cognitive Science **33**, 709–738 (2009)
38. Goldberg, M., Perfetti, C., Schneider, W.: Perceptual knowledge retrieval activates sensory brain regions. Journal of Neuroscience **26**, 4917–4921 (2006)
39. Gottlieb, J.: From thought to action: The parietal cortex as a bridge between perception, action, and cognition. Neuron **53**, 9–16 (2007)
40. beim Graben, P., Gerth, S., Vasishth, S.: Towards dynamical models of language-related brain potentials. Cognitive Neurodynamics **2**, 229–255 (2008)
41. beim Graben, P., Pinotsis, D., Saddy, D., Potthast, R.: Language processing with dynamic fields. Cognitive Neurodynamics **2**, 79–88 (2008)
42. Grill-Spector, K., Malach, R.: The human visual cortex. Annual Reviews of Neuroscience **27**, 649–677 (2004)
43. Harnad, S.: The symbol grounding problem. Physica D **42**, 335–346 (1990)
44. Hauser, M., Chomsky, N., Fitch, W.: The faculty of language: What is it, who has it, and how did it evolve? Science **298**, 1569–1579 (2002)
45. Hayward, W., Tarr, M.: Spatial language and spatial representation. Cognition **55**, 39–84 (1995)
46. Huyck, C.: A psycholinguistic model of natural language parsing implemented in simulated neurons. Cognitive Neurodynamics (in press)
47. Jackendoff, R.: Foundations of language. Oxford University Press, New York (2002)
48. Johnson, J.: Moving to higher ground: The dynamic field theory and the dynamics of visual cognition. New Ideas in Psychology **26**, 227–251 (2008)
49. Johnson, J., Spencer, J.P., Luck, S.J., Schner, G.: A dynamic neural field model of visual working memory and change detection. Psychological Science (in press)
50. Keysar, B., Barr, D., Balin, J., Brauner, J.: Taking perspective in conversation: The role of mutual knowledge in comprehension. Psychological Science **11**, 32–38 (2000)

51. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics **43**, 59–69 (1982)
52. Landau, B., Jackendoff, R.: "what" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences **16**, 217–265 (1993)
53. Levinson, S.: Space in language and cognition: Explorations in cognitive diversity. Cambridge University Press, Cambridge (2003)
54. Li, P.: Lexical organization and competition in first and second languages: Computational and neural mechanisms. Cognitive Science **33**, 629–664 (2009)
55. Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J., Schöner, G.: Robotic demonstrations of flexible reference frame transformations in spatial language. Manuscript under review (2009)
56. Lipinski, J., Spencer, J.P., Samuelson, L.: It's in the eye of the beholder: Spatial language and spatial memory use the same perceptual reference frames. In: L.B. Smith, M. Gasser, K. Mix (eds.) The spatial foundations of language. Oxford University Press (in press)
57. Logan, G.: Spatial attention and the apprehension of spatial relations. Journal of Experimental Psychology: Human Perception & Performance **20**, 1015–1036 (1994)
58. Logan, G.D., Sadler, D.D.: A computational analysis of the apprehension of spatial relations. In: P. Bloom, M. Peterson, L. Nadel, M. Garrett (eds.) Language and Space (Language, Speech, and Communication), pp. 493–529. MIT Press, Cambridge, MA (1996)
59. Magnuson, J., Dixon, J., Tanenhaus, M., Aslin, R.: The dynamics of lexical competition during spoken word recognition. Cognitive Science **31**, 1–24 (2007)
60. Martin, A., Haxby, J., Lalonde, F., Wiggs, C., Ungerleider, L.G.: Discrete cortical regions associated with knowledge of color and knowledge of action. Science **270**, 102–105 (1995)
61. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press, Chicago (1992)
62. McNeill, D.: Aspects of aspect. Gesture **3**, 1–17 (2003)
63. Moratz, R., Tenbrink, T.: Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. Spatial Cognition and Computation **6**, 63–106 (2006)
64. O'Keefe, J.: Vector grammar, places, and the functional role of the spatial prepositions in english. In: E. van der Zee, J. Slack (eds.) Representing direction in language and space. Oxford University Press., Oxford (2003)
65. Pfeifer, R., Bongard, J.: How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press, Cambridge (2006)
66. Pinker, S., Jackendoff, R.: The faculty of language: what's special about it? Cognition **95**, 201–236 (2005)
67. Pouget, A., Deneve, S., Duhamel, J.R.: A computational perspective on the neural basis of multisensory spatial representation. Nature Reviews Neuroscience **3**, 741–747 (2002)
68. Pulvermuller, F.: A brain perspective on language mechanisms: from discrete neuronal ensembles to serial order. Progress in Neurobiology **67**(2), 85–111 (2002)
69. Regier, T., Carlson, L.: Grounding spatial language in perception: An empirical and computational investigation. Journal of Experimental Psychology: General **130**(2), 273–298 (2001)
70. Samuelson, L., Schutte, A.R., Horst, J.: The dynamic nature of knowledge: Insights from a dynamic field model of children's novel noun generalization task. Cognition **110**, 322–345 (2009)
71. Sandamirskaya, Y., Schöner, G.: Dynamic field theory and embodied communication. In: I. Wachsmuth, G. Knoblich (eds.) Modeling communication with robots and virtual humans, Lecture Notes in Artificial Intelligence, Vol. 4930. Springer (2006)
72. Sandamirskaya, Y., Schöner, G.: Dynamic field theory of sequential action: A model and its implementation. In: B. Scassellati, G. Deak (eds.) The Seventh International Conference on Development and Learning, pp. 133–138. Monterey, CA (2008)
73. Schneegans, S., Schöner, G.: 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. Elsevier Ltd. (2008)
74. Schober, M.: Spatial perspective-taking in conversation. Cognition **47**, 1–24 (1993)
75. Schöner, G.: Dynamical systems approaches to cognition. In: R. Sun (ed.) The Cambridge handbook of computational psychology, pp. 101–126. Cambridge University Press (2008)
76. Schutte, A.R., Spencer, J.P.: Tests of the dynamic field theory and the spatial precision hypothesis: Capturing a qualitative developmental transition in spatial working memory. Journal of Experimental Psychology: Human Perception and Performance (in press)
77. Schutte, A.R., Spencer, J.P., Schner, G.: Testing the dynamic field theory: Working memory for locations becomes more spatially precise over development. Child Development **74**(5), 1393–1417 (2003)
78. Simmering, V.S., Schutte, A.R., Spencer, J.P.: Generalizing the dynamic field theory of spatial cognition across real and developmental time scales. Brain Research **1202**, 68–86 (2008)
79. Spencer, J.P., Perone, S., Johnson, J.: The dynamic field theory and embodied cognitive dynamics. In: J.P. Spencer, M. Thomas, J. McClelland

(eds.) Toward a New Grand Theory of Development? Connectionism and Dynamic Systems Theory Re-Considered. Oxford University Press, New York (in press)
80. Spencer, J.P., Simmering, V.S., Schutte, A.R., Schöner, G.: What does theoretical neuroscience have to offer the study of behavioral development? insights from a dynamic field theory of spatial cognition. In: J.M. Plumert, J.P. Spencer (eds.) Emerging landscapes of mind: Mapping the nature of change in spatial cognition, pp. 320–361. Oxford University Press, Oxford (2007)
81. Spivey, M., Dale, R.: On the continuity of mind: Toward a dynamical account of cognition. In: B. Ross (ed.) The psychology of learning and motivation, vol. 45, pp. 87–142. Elsevier, San Diego, CA (2004)
82. Spivey, M., Dale, R.: Continuous dynamics in real-time cognition. Current Directions in Psychological Science **15**, 207–211 (2006)
83. Spivey, M., M.J., T., Eberhard, K., Tanenhaus, M.: Linguistically mediated visual search. Psychological Science **12**, 282–286 (2001)
84. Sporns, O.: Complex neural dynamics. In: V. Jirsa, J.A.S. Kelso (eds.) Coordination dynamics: Issues and trends, pp. 197–215. Springer-Verlag, Berlin (2004)
85. Suzuki, W., Matsumoto, K., Tanaka, K.: Neuronal Responses to Object Images in the Macaque Inferotemporal Cortex at Different Stimulus Discrimination Levels. Journal of Neuroscience **26**(41), 10,524–10,535 (2006)
86. Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., Sedivy, J.: Integration of visual and linguistic information in spoken language comprehension. Science **268**, 1632–1634 (1995)
87. Taylor, H., Tversky, B.: Spatial mental models derived from survey and route descriptions. Journal of Memory and Language **31**, 261–282 (1992)
88. Tettamanti, M., Buccino, G., Saccuman, M., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S., Perani, D.: Listening to action-related sentences activates fronto-parietal motor circuits. Journal of Cognitive Neuroscience **17**, 273–281 (2005)
89. Thelen, E., Schner, G., Scheier, C., Smith, L.B.: The dynamics of embodiment: A dynamic field theory of infant perseverative reaching errors. Behavioral and Brain Sciences **24**, 1–86 (2001)
90. Thelen, E., Smith, L.B.: A Dynamic Systems Approach to the Development of Cognition and Action. MIT Press, Cambridge (1994)
91. Tononi, G., Edelman, G.M., Sporns, O.: Complexity and coherency: Integrating information in the brain. TRENDS in Cognitive Sciences **2**, 474–484 (1998)
92. Wallentin, M., Ostergaard, S., Lund, T., Ostergaard, L., Roepstorff, A.: Concrete spatil language: See what i mean? Brain and Language **92**, 221–233 (2005)
93. Wersing, H., Steil, J.J., Ritter, H.: A Competitive-Layer Model for Feature Binding and Sensory Segmentation. Neural Comp. **13**(2), 357–387 (2001)
94. Wilimzig, C., Schneider, S., Schner, G.: The time course of saccadic decision making: Dynamic field theory. Neural Networks **19**, 1059–1074 (2006)
95. Wilimzig, C., Schöner, G.: How categorical behavior emerges from continuous neural representations: Dynamic field theory (in preparation)
96. Wilson, H.R., Cowan, J.D.: A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Kybernetik **13**, 55–80 (1973)
97. Xing, J., Anderson, R.A.: Models of the posterior parietal cortex which perform multimodal integration and represent space in several coordinate frames. Journal of Cognitive Neuroscience **12**(4), 601–614 (2000)

7 Acknowledgements

The authors wish to thank Vanessa Simmering and Christian Faubel for helpful comments on a previous version of this manuscript.