An Embodied Account of Spatial Language Grounding

Inauguraldissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

in der

Fakultät für Psychologie

der

Ruhr-Universität Bochum

vorgelegt von

Jonas Lins

Bochum, März 2018

Referent: Prof. Dr. Gregor Schöner
Korreferent: Prof. Dr. Sen Cheng

Tag der mündlichen Prüfung: _____

Abstract

Theories of embodied cognition hold that cognition is tightly linked to processes in sensorimotor systems. Explaining how language connects to modal representations within these systems poses a challenge, since the discrete, sequential format in which language conveys information is fundamentally different from that of continuous modal representations. Existing evidence for a link between these two representational realms is typically based on coactivation effects, or on perceptual or motor performance being facilitated or impeded by concurrent language input. The current thesis complements this research with a different type of evidence that focuses more specifically on the processes which ground linguistic input in sensorimotor spaces. Language about spatial relations was used as a test case.

First, a neural process model of the mechanisms linking spatial language to modal representations was developed. The model is based on the theoretical framework of Dynamic Field Theory, which describes neural activation patterns at the population level. Dynamic neural fields and discrete nodes were combined into a seamless neural architecture that autonomously grounds spatial phrases in vision. The model proposes a prototype mechanism for grounding, which was used as a heuristic to derive and interpret potential effects in subsequent behavioral experiments.

Second, a novel computer mouse tracking paradigm was devised and used in seven experiments to measure behavioral signatures of grounding processes. Participants saw a spatial phrase such as "The green item to the left of the red one", followed by an array of colored items, and moved the mouse cursor onto the target item of the spatial phrase (here, the green one). Visual items implicated in the process of grounding were found to affect mouse movement trajectories. Most importantly, trajectories were attracted by the reference item of the phrase (here, red), by distractors sharing the target color (here, green), and by a competing relational pair. These effects were interpreted as showing attentional allocation to these items in the course of grounding. A bias into the direction described by the spatial term (here, "left of") was more akin to classical embodiment effects.

The experiments allowed to observe the online specification of response movements and its modulation by processes of language grounding operating on sensorimotor substrates. This supports that, in this scenario, language is grounded in sensorimotor systems rather than being processed on an abstract cognitive level. In summary, this thesis provides a novel type of evidence for the embodiment of language understanding, by leveraging the close linkage of perception and action systems within the sensory-motor loop.

Short Contents

Contents

List of Tables

List of Figures

List of Abbreviations

ANOVA Analysis of variance

CoD Condition of dissatisfaction

CoM Center of mass

CoS Condition of satisfaction

DF Dynamic neural field

DFT Dynamic Field Theory

DPA Distribution of population activation

IoR Inhibition of return

SD Standard deviation

Chapter 1

Introduction

Imagine having dinner at a friend's house. The table is cluttered with plates, platters, glasses, cutlery, napkins, and bottles. "Would you pass me the salt, please?", one of your friends asks. The word forms in this request have no inherent connection to the things in your visual surroundings, and the order of sentence components does not relate in any direct way to the physical events it aims to bring about.

Yet you are able to decode the phrase's syntax, understand what motor action is required and, most importantly, link the word "salt" to a bundle of perceptual features located at a specific position in your environment. You could have done this even if your friend had spoken French (provided you know French), "Tu me passes le sel, s'il te plaît?" — even though the words and grammar are wildly different, they refer to the same entities in the world.

Having singled out the salt shaker, you start to reach toward its position, but you discover that there are two alternatives — the salt and pepper shaker look alike! "It's the shaker to the left of the wine bottle!", your friend helps. Again, you map from auditory to visual patterns, but this time the mapping requires even more complex processes, because the spatial term "left of" refers to a higher order relationship that is not explicitly specified in the visual image on your retinas.

Instead, you have to localize multiple feature bundles that correspond to "wine bottle" and "shaker" and assess which configuration of eligible items matches the pattern implied by the spatial term. You master this effortlessly as well, single out one of the shakers, adjust reach parameters, and pass your friend the salt.

How does human cognition master the feats evident in this example? How is language grounded in the world despite its constituents being structurally different from perception and motor action? The current thesis elaborates this

question under the premise of embodied cognition, which emphasizes the continuous coupling of cognition, perception, and action. Spatial relational language serves as a test case that can be sufficiently simplified (as opposed to, for instance, analogies or metaphor) while it already poses an abstraction from immediate sensory input. How spatial relations are grounded is examined by means of a neural dynamic process model and an experimental assessment of behavioral signatures of the grounding processes' embodied nature.

1.1 Linking language to embodied cognition

The initial example illustrates two fundamental capabilities that humans have: Comprehending abstract, amodal, discrete symbols, here supplied in the form of language, and interacting through perception and motor action with an environment that is continuous and extended in space and time.

How arbitrary symbols that are unrelated in shape to what they stand for in the world can be mapped to their concrete referents has come to be known as the symbol grounding problem (Harnad, 1990; Searle, 1980). The recognition of symbol grounding as a problem was originally a reaction to the classical view that cognition can be understood as a formal, rule-based system operating on abstract symbols, independent of the sensorimotor interfaces and neural implementation (Pylyshyn, 1980; Fodor, 1983).

The fundamental criticism was that such accounts cannot explain how abstract symbols acquire semantic content; the conclusion was that such grounding could only be achieved through a bottom-up approach, starting from the projections of distal objects on the sensory surfaces and from there proceeding to categorization and identification (Harnad, 1990; Barsalou, 1999; Glenberg, 1997).

In line with these concerns, the last two decades have seen the symbolic account challenged by the idea of embodied cognition (M. Wilson, 2002; Fahim & Rezanejad, 2014; Schöner, 2008; Schneegans & Schöner, 2008; Meteyard et al., 2012) and the closely associated notion of grounded cognition (Barsalou, 2008, 1999, 2010; Glenberg, 1997; Lakoff & Johnson, 1999).

Embodied cognition emphasizes that the physical make up of the body and its constant interaction with a richly structured environment shape and have shaped cognition decisively during evolution, development, and learning (M. Wilson, 2002; Schöner, 2008; Fahim & Rezanejad, 2014). It therefore holds that, to understand cognition, the structure of sensory and motor systems must be taken into account, as well as that of other neural substrates.

Embodied cognition furthermore emphasizes that cognition is situated, meaning that the nervous system is at all times linked to the world outside and inside the body through its sensorimotor surfaces. This continuous linkage imposes specific requirements on the processes and neural mechanisms that constitute cognition, which may be neglected when overly abstracting from embodiment (Schöner, 2008). The current thesis commits to a rigorous variant of the embodiment stance, which contends that the neural processes and substrates for perception, action, and cognition all share certain fundamental properties (Schöner, 2008; Schöner et al., 2015). The latter two points will be covered in greater detail in Section 1.5.

Grounded cognition has strong overlap with embodied cognition, both in the underlying postulates and with respect to the community of proponents, and is often not clearly separated from it in the literature. It is perhaps best described as extending the idea of embodiment to 'offline' cognition about absent things or situations. While highlighting that human cognition may *also* take place somewhat independently of the immediate situational context, it emphasizes that it is even then grounded in sensorimotor experience. This grounding in sensorimotor experience is thought to occur as a partial re-enactment and recombination of previously experienced sensorimotor activation patterns within modality-specific substrates (Barsalou, 2008, 1999; Lakoff & Johnson, 1999; Glenberg, 1997; Meyer & Damasio, 2009).

Embodied and grounded cognition run counter to the notion that cognition can be explained as purely abstract manipulation of amodal symbols. At first glance, symbol grounding seizes to be a problem under this premise: Doing away with amodal symbols also eliminates the need for transduction between two otherwise disconnected representational realms. If the neural mechanisms that create cognition are similar to or congruent with those of perception and motor action, then mapping between them becomes trivial.

However, this is not the whole truth. Although it is easier to conceive how sensorimotor systems and cognition are connected under the framework of embodied cognition, complex linking problems remain to be solved. In general, a complementary role for amodal representations and processes is not ruled out in embodied and grounded cognition (e.g., Zwaan, 2014; Barsalou et al., 2008; Dove, 2009; for review, see Meteyard et al., 2012). However, language, with its obvious distinctness from grounded representations (Glenberg, 1997), is probably the most obvious example for a linking problem in human cognition that is reminiscent of symbol grounding.

While language is, of course, perceived through the senses like any other

stimulus, it is 'amodal' in the sense that the discrete representational format which it uses to convey content is fundamentally different from the continuous nature of sensorimotor experience. For most of language, the forms of words bear no structural relationship to their meanings (e.g., Monaghan et al., 2014). The phonological form of "red", for instance, is not more similar to "orange" than to "green", although the color red is perceived as more similar to orange than to green. Yet, this system of arbitrary forms has powerful combinatorial properties, such as the capacity to create an infinite number of different combinations from a finite number of atomic elements (Chomsky, 2005; e.g., combining words into ever new sentences and being able to understand such novel combinations).

Thus, language seems to be quite different from perception and action, and yet has the power to conjure up complex and novel combinations of percepts, actions, or cognitive states, which proponents of embodied cognition believe to be realized as grounded patterns of activation. This means that even though the space of linguistic forms does not preserve the structure of the grounded spaces of sensorimotor parameters (Gasser, 2004), the human nervous system has developed mechanisms which map between the two. In summary, while the symbol grounding problem in its strong form vanishes when cognition is embodied, the question arises even more explicitly how language couples into the embodied cognitive system.

Here, a proposal is made how this coupling may occur, in the form of a neural process model of spatial language grounding, and behavioral experiments are described that confirm the hypothesized mechanisms. The next section describes evidence which suggests that there is indeed a close association between language understanding and sensorimotor systems, illustrating the nature of current experimental research in the area. After that, the goal and contribution of the current thesis are described, followed by its theoretical vantage point and aspects that are relevant to the specific approach taken.

1.2 Evidence for the embodiment of language

Multiple forms of evidence suggest that language understanding is embodied in terms of engaging substrates that are also implicated in perceiving or acting.

Imaging techniques show this most directly, as an overlap of brain areas activated during language comprehension with those activated during actual perception or motor action. For instance, listening to action-related sentences

leads to activation in motor cortex, as evidenced by a modulation of motor-evoked potentials in the body part that the described action would engage (hand or foot; Buccino et al., 2005). Hearing sentences that describe actions such as "I grasp a knife", activates a network of frontal, temporal, and parietal regions which overlaps with regions that are active during execution and observation of actions (Tettamanti et al., 2005). Activation of motor areas when hearing action words is somatotopically related to the involved body parts (Pulvermüller et al., 2005).

For the auditory modality, Kiefer et al. (2008) report that visually presented words describing objects that are associated with acoustic features, such as "telephone", activate temporal areas that are also active when sound is perceived. Similar evidence exists for gustatory words, which recruit primary and secondary gustatory cortices (Barrós-Loscertales et al., 2012). Evidence with respect to visual areas is more sparse, but is has been shown, for instance, that sentences describing visual events (such as reading a book) activate parts of the secondary visual cortex (Desai et al., 2010).

A common approach of behavioral studies aims to show that mental simulation takes place during language understanding through demonstrating that the contents of the purported simulations affect other, non-linguistic tasks.

In an approach pioneered by Stanfield & Zwaan (2001), participants read a sentence that includes an object, are then presented with a picture of an object, and have to indicate whether the picture matches the object from the sentence. In match trials, a visual property of the depicted object is manipulated to be either consistent or inconsistent with a property implied by the sentence. An impact of sentence-picture consistency on response latency or other behavioral measures is interpreted as shared sensorimotor substrates being engaged by language and visual stimuli. According to proponents of grounded cognition, such associations should not occur if language is processed in an amodal way (Barsalou, 1999).

Applying this experimental approach, Zwaan et al. (2002) used sentences that implied a specific object shape, such as "The ranger saw the eagle in the sky", implying an eagle with outstretched wings. The subsequent picture would show either an eagle in its nest (folded wings) or in the sky (outstretched wings). Recognition and naming latencies were shorter if the shape in the picture was consistent with that implied in the sentence. Similar evidence exists for other visual features, including orientation (Stanfield & Zwaan, 2001), color (Zwaan & Pecher, 2012; see also T. Richter & Zwaan, 2009), degree of visibility (Yaxley & Zwaan, 2007), and object motion (Zwaan

et al., 2004; for review, see Zwaan & Pecher, 2012).

Conversely, visual stimulation affects language processing. For instance, when viewing motion stimuli (e.g., spirals moving in- or outward) while simultaneously hearing sentences that imply motion (e.g., "The car approached you."), sentence sensibility judgments are impeded when motion directions match (Kaschak et al., 2005). The authors interpret this as interference in sensorimotor substrates shared between language understanding and visual motion processing.

That language understanding is also linked to activation of motor systems has been shown using a similar logic. Glenberg & Kaschak (2002; see also, Kaschak & Borreggine, 2008, for an exploration of the effect's time course) let their participants read and judge sentences for sensibility that were either nonsensical, implied motion toward the own body (e.g., "Andy delivered the pizza to you."), or motion away from it (e.g., "You delivered the pizza to Andy."). Speeded judgments had to be made by pressing one of two buttons, where one button position required to move the hand away from the body, while the other required to move the hand closer to the body. Reaction latencies where shorter when the required direction of hand movement was consistent with the direction of motion implied in the sentence. This extended to sentences describing concrete object transfers, as well as imperatives, and to more abstract sentences (e.g., "Liz told you the story.").

A similar experiment (Sell & Kaschak, 2011) revealed that an effect of compatibility between linguistic input and motor action also occurs when participants understand sentences about future versus past events. Sentences describing future events were associated with faster responses when the hand had to be moved away from the body (i.e., forward), and vice versa, but not when the hands were stationary on the response buttons. This suggests that even language about abstract concepts may be grounded in sensorimotor systems.

Richardson et al. (2003; see also, Bergen et al., 2007) as well report evidence consistent with this idea, for the domains of perception and memory. Their participants heard sentences that included concrete or abstract verbs commonly associated with either vertical or horizontal directionality. (This was determined in a norming task by Richardson et al., 2001; "respect", for instance, was associated with a vertical, "argue with" with a horizontal directionality.) The sentences were presented simultaneously with a visual categorization or a visual memory task. It was found that task performance was affected by verb directionality. Briefly flashed visual targets were identified

faster and more reliably if they were presented in a location inconsistent with the directionality of the verb (e.g., when the visual stimulus was located in the upper or lower screen center and the verb implied a horizontal directionality). Conversely, images illustrating the presented sentence were recalled better when they were arranged in an orientation consistent with the verb's implied directionality.

Transcending the above effects of interference and facilitation, language may also bias movement parameters in a manner consistent with its semantic content. Glover & Dixon (2002) printed the words "LARGE" or "SMALL" on equally sized grasping targets, instructing participants to ignore the words, and found that initial grip aperture during grasping was wider or narrower, consistent with the word on the target. Gentilucci et al. (2000) similarly report that movement parameters such as peak arm velocity and acceleration were affected by the words "NEAR" and "FAR" printed on grasped objects.

Zwaan et al. (2012) had their participants judge the sensibility of visually presented sentences that implied forward or backward movement, for instance, "The knight bowed to the king". Responses had to be indicated by leaning left or right. While doing so, participants also shifted their postural balance forward or backward, in line with the implied direction.

An experiment by Tower-Richardi et al. (2012) shows a similar impact of language on motor action. This study is particularly relevant for the current work since both the method and the observed effect overlap with aspects of the experiments reported in Chapter 3. In the study, participants used a computer mouse to move a cursor from the screen center to one of four target boxes arranged around it. In each trial, the target was indicated by a centrally presented word reading either "UP", "DOWN", "LEFT", or "RIGHT" (in another condition that yielded similar results response direction was indicated by arrows). Critically, a masked prime was flashed in the beginning of each trial, which participants did not consciously perceive. The primes were either non-words or one of the words "NORTH", "SOUTH", "EAST", and "WEST". It was found that the directional primes, but not non-words, biased mouse trajectories into a direction consistent with the prime. For instance, participants made a slight detour to the right of the straight path to the target when they had to move upward but were primed with the word "EAST". This hints at spatial concepts being grounded in sensorimotor systems (Tower-Richardi et al., 2012).

Experimental approaches as those described above are abundant in the literature, mostly yielding results which similarly suggest that language com-

prehension is correlated with activation of modal systems (for further review, see Barsalou, 2008; Pecher et al., 2011; Zwaan, 2014).

1.3 Critique of the embodiment stance

Findings as those reviewed above are commonly deemed supportive of the claim that language understanding and other forms of higher-order cognition depend on sensory and motor systems.

It has been argued, however, that the evidence is largely correlational and unspecific in nature (Willems & Francken, 2012; Dove, 2009). The question is therefore open whether activation in sensorimotor regions is causally implicated in language comprehension and conceptual processing, or whether it represents an epiphenomenon of abstract symbol processing (Mahon, 2015; Dove, 2009).

A related shortcoming in the field of embodied cognition is the wide variety of theoretical positions, often vaguely specified, about the degree to which different cognitive domains are embodied, partly embodied, or reliant on interaction with amodal components (but see Meteyard et al., 2012, for an attempt of consolidation). The vagueness of the theoretical claims makes them difficult to transform into formal models and severely restricts their utility in guiding and interpreting research (Goldinger et al., 2016).

In line with this, it has been suggested that research in embodied cognition has reached a stage where it should focus more strongly on *how* and *when*, that is, under which circumstances, language and conceptual processing are grounded in sensorimotor systems, for instance, through computational modeling (Willems & Francken, 2012).

The author of the current work similarly contends that the available evidence is too easily interpretable in favor of differing theoretical positions, owed to both the correlational nature of the evidence and vaguely formulated theories that remain largely verbal. From the perspective of a researcher used to devising neural processes models, the lack of theoretical specificity may be due to a lack of concomitant investigations how the required mechanisms may pan out on the level of neural processes. Through the added constraints arising from the restricted toolbox that biological nervous systems provide, such accounts can uncover issues of theoretical views that would otherwise go unnoticed. Furthermore, through forcing concrete realizations of verbal theories or their components, they may provide a heuristics for developing empirical approaches to differentiate between theories.

1.4 Thesis goal and outline

The current thesis approaches the topic of embodied cognition in concrete terms. A specific scenario is examined that brings together linguistic, sensory, and motor aspects. In the scenario, phrases about spatial relations, such as "The green item to the left of the red one" must be linked to or 'grounded in' a visual scene. Grounding is in this context understood as finding the referents of the spatial phrase in the visual environment through controlled interaction between discrete, language-like representations and sensorimotor substrates.

The approach taken is based on two trivially true assertions. First, language is, at some level, different in format from representations on the sensorimotor level. Second, humans are able to link language to entities in the current sensorimotor environment and vice versa (enabling language understanding and production). It follows from these assertions that neural mechanisms for grounding must exist. In other words, the mere fact that language *can* be linked to the sensed environment is evidence for neural processes that connect discrete to sensorimotor representations.

Explaining how this occurs already provides a difficult linking problem to solve, independent of whether or not one is mandatory for the other. Therefore, possible mechanisms of embodiment are here explored not by considering how language is understood in the absence of what it refers to, as is commonly done in grounded cognition research, but by investigating how language is neurally linked to the current sensory environment.

This is done in two ways. First, a neural process model is presented which implements a prototypical grounding mechanism for the exemplar case of spatial language. Second, a set of experiments is described that probe behavioral signatures of the component processes implicated in the grounding of spatial language and thereby demonstrate that they engage sensorimotor substrates.

Both the model and the experimental paradigm focus on investigating the specific processes that may underlie the grounding of language. They consider not mere interference or facilitation, nor a diffuse influence of language embodiment on motor action, but the fleeting processing steps which unfold in time during grounding a relational phrase.

The manner in which the underlying cognitive processes are embedded in the sensory-motor loop is likewise made concrete, by building on a theoretical framework that formalizes the properties of neural substrates that allow the

required balance of stability and flexibility in the face of a constant linkage to sensorimotor surfaces.

Spatial relations are chosen as a test case because they pose a conceptual class that refers to something which is derived from a state of affairs in the world, without themselves being explicitly available on the sensory surfaces. They thereby illustrate the cognitive capacity of forming a semantic representation of abstract nature.

The neural process model will be described in Chapter 2, and the behavioral experiments in Chapter 3. Prior to this, a number of topics will be covered for preparation. Section 1.5 describes the specific stance of embodiment that underlies both the model and the experiments. It is closely associated with the theoretical framework of Dynamic Field Theory and formalizes the embeddedness of cognition within situated cognitive agents. Section 1.6 describes the experimental method of computer mouse tracking, which in recent years has been employed to obtain rich behavioral measures of cognitive processes, and which is applied in a novel form here. Section 1.7 provides an overview of the cognitive processes implicated in evaluating visual spatial relations, which is relevant for both the model and the experiments. Finally, Section 1.8 describes Dynamic Field Theory in detail, to provide the theoretical language in which the neural process model is formulated.

1.5 Embodiment in Dynamic Field Theory

Dynamic Field Theory (DFT) provides a formal account of embodied cognition which captures the seamless coupling of perception, cognition, and action in a concrete mathematical framework (Schöner, 2008; Schöner et al., 2015; Schneegans & Schöner, 2008; Lins & Schöner, 2014). In this framework, neural interaction adds a layer of stability to embodied cognitive systems that to a degree decouples their behavior from the immediate sensory environment. The simple behavioral dynamics of a situated embodied system that lacks this type of stability will be considered first, to then demonstrate how adding neural interaction may enhance behavioral flexibility.

To illustrate behavioral dynamics consider a Braitenberg vehicle (Braitenberg, 1984; Schöner, Faubel, et al., 2015). In such a vehicle, sensors are coupled to effectors in a feed-forward manner, so that when a sensor is excited effector activity is directly affected. In the vehicle shown in Figure 1.1, each of two light sensors mounted in the front of the vehicle is coupled to the ipsilateral wheel at the vehicle's back. The coupling is inhibitory, so that the turning

rate of a wheel is reduced when the connected sensor is activated (baseline wheel turning rate is assumed to be greater than zero). Thus, when a light is located to the left of the vehicle, activating the left sensor more strongly than the right one, the left wheel's turning rate is reduced more than that of the right wheel. In consequence, the vehicle turns to the left, towards the light. As the vehicle turns, the sensors become more equally stimulated, so that wheel turning rate becomes more equal as well. The vehicle thus turns until it faces the light, moving toward it until sensor activation becomes so strong that the wheels stop turning.

How the turning behavior unfolds in time under this particular setup can be captured by a simple dynamics of vehicle turning rate defined over heading direction, driven by the distribution of light intensity relative to the vehicle's heading. In the case of a single light stimulus as described so far this behavioral dynamics has a single attractor at the angular position of the light source (for a more detailed description of this dynamics as well as the following one please refer to Schöner, Faubel, et al., 2015).

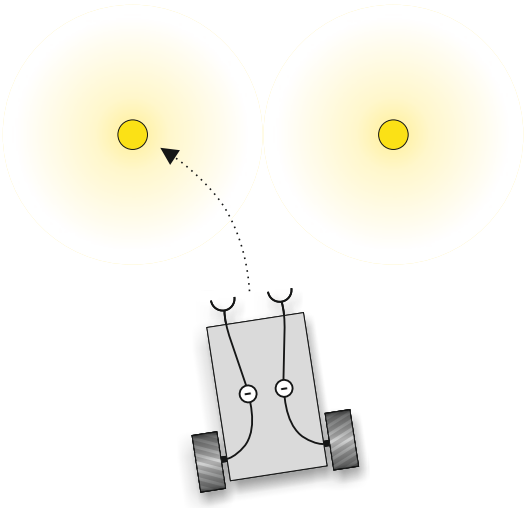


Figure 1.1: Behavior of a Braitenberg vehicle with each light sensor coupled inhibitorily to the ipsilateral wheel motor, in an environment with two light sources.

When there are two light sources present in the environment, for instance, to the left and right of the vehicle as shown in Figure 1.1, two attractors emerge located at the angular positions of the two lights. To which light the vehicle will turn is then determined by its initial heading direction, that is, which basin of attraction the system resided in when the two lights appeared. Turning to one stimulus is a selection decision insofar as the vehicle picks one identical stimulus over the other, setting its behavioral goal such that the ignored stimulus has little impact on the final outcome.

The capacity to make a selection decision arises not from the feed-forward connections between sensors and effectors; these merely map a given input to the same output each time this input occurs. It only arises by situating the system and its body in a structured environment. None of the components alone, neither the body shape, nor the connection scheme, nor the environment, can

be identified as enabling the selection decision, but it instead emerges from the sensory-motor *loop* formed by body, feed-forward connections, and the environment. This is a fundamental way in which embodiment shapes behavior.

As a consequence of this tight coupling, however, the selected goal will only be pursued as long as no external perturbation occurs that pushes the system into a different basin of attraction, and as long as the target stimulus is present on the sensory surface. When the targeted stimulus vanishes, for instance, due to an obstacle occluding it, the vehicle will abandon its previous goal and move toward the other stimulus. The stability of the decision is still entirely dependent on the stimulus landscape.

The system thus lacks a hallmark of biological cognition: the capability to shield an established decision from being overwritten by sensory input. A human on a nightly hike, having decided to ignore the close village lights and instead move toward a distant lit up cottage, will pursue that goal even when the cottage lights are momentarily occluded by trees. In fact, the tendency to stick to decisions somewhat independent of stimulation is present already in seemingly low-level cases of perception (e.g., Hock et al., 1997; Chambers & Pressnitzer, 2014). To achieve this increased behavioral flexibility, a layer of stability is required that breaks the feed-forward coupling between environment and behavior.

Within the framework of DFT, this layer is realized by dynamic neural fields (DFs; the term *field* is used interchangeably here). A detailed explanation of the DFs and their grounding in neurophysiology will be provided in Section 1.8, while here only aspects relevant for the current context are covered. Each DF is modeled by a differential equation which describes a continuous distribution of activation over a population of neurons and its evolution in time. The neurons captured by the field are assumed to be broadly tuned to different preferred values along a sensory, motor, or cognitive dimension, for instance, retinal space, color, or movement direction. Thus, when a localized input is present along the dimension spanned by a DF, such as a visual item at a particular position, the field site corresponding to the respective value is elevated above resting level, forming an input-driven hill of activation.

When field activation crosses a soft output threshold above the resting level, localized output is generated. Output may, on the one hand, impact downstream substrates. On the other hand, it drives lateral interaction within the DF which is mediated by recurrent connectivity between neighboring field sites and takes the form of local excitation and surround inhibition. With

the onset of lateral interaction, field activation is no longer purely input-driven. Local excitation may amplify even near-threshold input to stable peaks. Strong local excitatory connectivity may sustain peaks after the input has seized. Inhibitory interaction may restrict the number of peaks that can coexist in a DF, forcing selection when multiple localized inputs arrive.

The formation, selection, and suppression of peaks are decisions similar in effect to the selection realized by the behavioral dynamics of the vehicle described above. In DFs, however, the decisions are based on neural dynamics. The recurrent neural connectivity breaks the fixed feed-forward coupling between input and field activation. This feature of DFs allows for stable neural representations and shields inner states against input changes and external perturbations. In turn, a field connecting sensory and motor systems may break the feed-forward coupling between environment and behavior.

The decoupling is not absolute, though. Supra-threshold activation in DFs is not entirely insensitive to changes in the distribution of inputs, which still impacts field activation as before. A moving localized input, for instance, may be tracked by a peak (Faubel & Zibner, 2010; Bicho et al., 2000). A new, sufficiently strong localized input may lead to the decay of an existing peak through inhibition. This allows updating, for instance, to re-align visual working memory with a scene after a saccade (Johnson, Spencer, & Schöner, 2009). The exact degree of decoupling is dependent on the strength of lateral interaction and may be adjusted to the demands of the particular role a field serves in a cognitive system.

In the simple picture of the Braitenberg vehicle, adding a DF between sensors and motors allows the vehicle to select and pursue goals even when the original stimuli are momentarily occluded or vanish completely. Such a field would be defined over heading direction, receive input from the vehicle's sensors in a point-spread manner, and send its output to the motors. From the perspective of the motor surface (i.e., the wheels) the field would assume the role of the distribution of light in the environment that originally drove the behavioral dynamics. From the view of dynamics, a peak in the field effectively sets an attractor in the space of heading direction by influencing turning rate through its impact on each wheel's turning rate.

With the field in place, decisions about light sources being present or not, and which to pursue, are made and stabilized within the field and not only by where the vehicle faces. A detailed account of such a system can be found in Schöner, Faubel, et al. (2015), including a review of a robotic implementation demonstrating the rich capabilities emerging from a similar setup (Bicho et

al., 2000; there, an array of five sensors was used to provide more precisely localized input to the field). Schöner, Faubel, et al. (2015) also outline that the principles of controlling motor action in this way carry over to much more complex, biologically realistic scenarios, such as muscle control for limb movement.

In sum, DFs add inner stability to situated cognitive systems while staying true to the tenets of embodied cognition. No concepts of discontinuity and no non-neural steps need to be invoked when embedding dynamic fields into an embodied system. No transduction between representational formats is required, nor must a final decision be 'read out' at any point to drive a behavioral response or judge model performance. While the flow of activation through the sensory-motor loop is fundamentally continuous, dynamic instabilities that arise from lateral interaction explain how cognitive processing may decouple from the environmental context. These principles also serve to couple and decouple neural fields from each other. As will later be shown (see Chapter 2), this may be neurally organized to occur in a controlled manner, such that computational steps emerge from continuous dynamical systems architectures (M. Richter et al., 2012; Sandamirskaya & Schöner, 2010).

DFT thus allows to capture perceptual, cognitive, and motor tasks by a unified set of concepts. This is possible since each of these domains is ultimately part of the sensorimotor loop and therefore subject to the same requirements. As a result it is usually not possible to unambiguously assign a neural field to a specific domain. This can again be made concrete with the above example of a vehicle whose sensors and motors are connected via a DF (a rough conceptual illustration is provided in the left part of Figure 1.2). As described, the field may create a peak of activation at the position of an environmental stimulus. The localized input that initially brings activation above threshold, and ultimately leads to the dynamic instability underlying peak formation, comes from the sensors of the vehicle. The motor command that the wheels receive, on the other hand, is also based on the field output. Thus, the decision occurring in the field is both perceptual and motoric in nature. The distinction between perception, action, and cognition that is classically made in cognitive psychology is thus an arbitrary one from the perspective of DFT.

The postulate of DFT adopted here is that this idea transfers to more complex scenarios, such as the grounding of spatial language that is used as a test case in the current work (right part of Figure 1.2). In this case as well, decisions made in the neural substrate of the DFs cannot be clearly assigned to sensory, cognitive, and motor categories. Peaks in the field architecture

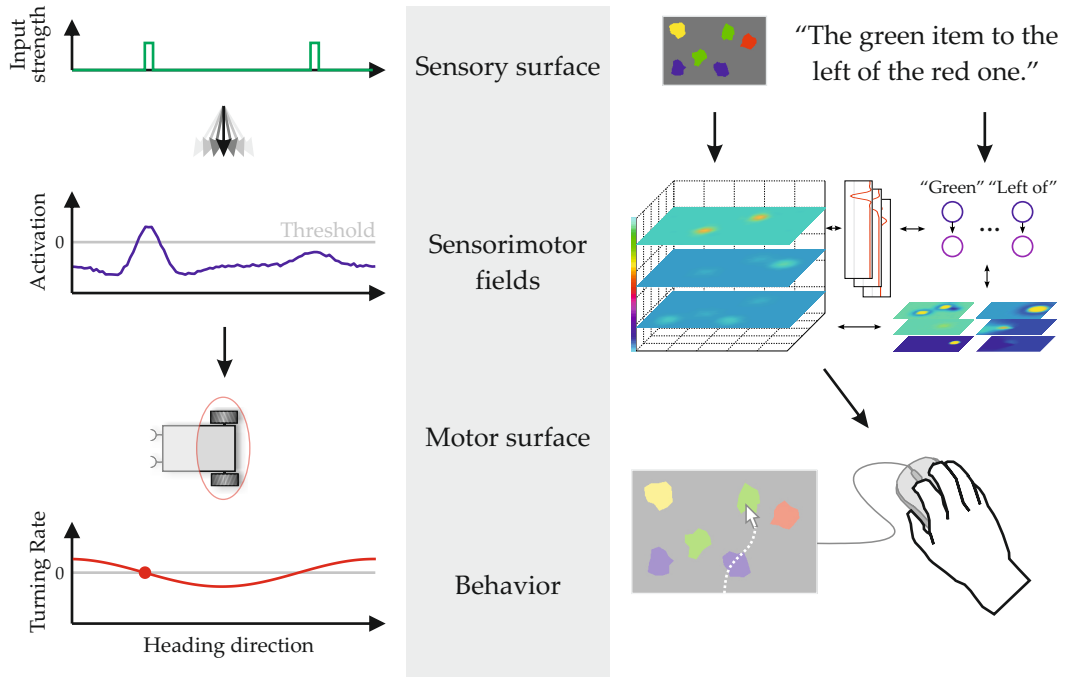


Figure 1.2: A conceptual illustration how dynamic fields couple continuously into the sensory-motor loop and thus unify perceptual, cognitive, and behavioral aspects. The left part illustrates the simple scenario discussed in the text where a dynamic field makes a decision about which stimulus to pursue (the bottom plot shows the behavioral dynamics arising from this setup, with the marked zero crossing representing an attractor). The right part transfers this picture to the more complex scenario of spatial language grounding and computer mouse tracking covered in this thesis. Fields of different dimensionality that are part of the model are shown in the center plot on the right but do not represent the full model architecture.

represent sensory decisions about stimuli being present. At the same time, however, attentional biases that arise from saliency and language input make the waxing and waning of these peaks cognitive decisions about their current relevance for the cognitive task. The appropriate evolution of these same decisions ultimately leads to 'understanding' of the relational phrase and may directly inform behavior, such as selecting the correct visual item with a computer mouse, as probed in the experiments presented later.

Cisek (2007; see also, Cisek & Kalaska, 2010) makes a very similar proposal. Based on a review of research on visually guided action he concludes that activation in sensorimotor maps along the dorsal visual stream does not typically capture the world per se, that is, independent of function, but rather specifies multiple potential action targets that simultaneously compete for execution within the very same substrates. Further decision variables are hypothesized to bias this competition through reciprocal connectivity with other regions, such as the ventral visual pathway, prefrontal cortex, and the

basal ganglia. At the basis of this proposal lies the view that the evolution of the central nervous system occurred primarily under the pressure to guide real-time action, while advanced cognition has developed only based on the mechanisms serving this primal task (Cisek, 2007).

Further evidence consistent with this view is based on reaching and computer mouse tracking and similarly suggests a close interplay of cognition, action, and perception; this evidence will be discussed in the following section.

In line with the framework sketched in this section, the current thesis is based on the hypothesis that even the particularly human capability of language understanding is couched in the embodied loop of sensing and acting, and is therefore subject to the same requirements of coupling and decoupling from the world in a controlled way. This is instantiated by the model presented in Chapter 2 for the task of grounding spatial relations in visual scenes. Note in this context that to remain consistent with the embodiment stance, process models within the framework of DFT do not need to be explicitly embodied (e.g., through robotic implementation). It is sufficient that the principles that underlie DFT are heeded so that the possibility of coupling a model to real sensory and motor systems to generate stable behavior is retained (Schöner, 2008). On the other hand, the linkage of sensory, cognitive, and motor processes is leveraged in a set of experiments, presented in Chapter 3, which aim to show behavioral signatures of language grounding to confirm its embodied nature.

1.6 Probing embodiment with computer mouse tracking

The previous section outlined the view that all cognition is tightly coupled in the sensory-motor loop and shares neural substrates and structural properties with perception and action. Under this premise, it can be expected that ongoing motor action reflects the evolution of concurrent cognitive processes in a manner that transcends facilitation or interference.

A prerequisite is that movement plans evolve continuously over time, rather than emerge abruptly. That this is indeed the case has been shown in the timed movement initiation paradigm by Ghez et al. (1997). Their participants made hand movements to visual targets that were situated in two different directions around the hand. Movement initiation time was prescribed

by the fourth tone of a metronome and therefore known in advance. The final target position was cued only shortly before movement initiation, thereby allowing variable amounts of time for movement planning (0–400 milliseconds).

Ghez et al. found that shorter planning intervals lead to movements into a default direction between the potential targets. Critically, however, movement directions became progressively more congruent with the actual target direction with increasing interval duration; first adjustments toward target direction were observable at intervals of approximately 100 milliseconds and precision increased up to 300 milliseconds. This was modulated by the angular separation of the targets: if it was greater than 60 degrees, participants tended to choose one of the two directions at random rather than moving into intermediate directions.

These results show that the process which specifies movement direction is continuous in time and affected by task parameters (Erlhagen & Schöner, 2002). This is complemented by neural data which shows that multiple possible movements may be specified simultaneously by neuron populations in premotor cortex when prior information about upcoming movements is ambiguous (e.g., Cisek & Kalaska, 2005; Bastian et al., 1998). In line with this, recent data shows that pointing trajectories are biased according to the spatial distribution of potential target locations when the final target is indicated only after movement onset (Gallivan & Chapman, 2014; Chapman et al., 2010).

In the above experiments, uncertainty over the final target location was induced by previewing multiple potential targets, and target selection was enabled by simple perceptual cues. The experimental technique of computer mouse tracking (Spivey et al., 2005; for review, see Freeman et al., 2011) and similar approaches (for review, see Song & Nakayama, 2009) have moved beyond this, by showing that other task variables may similarly modulate target uncertainty and that this affects trajectories accordingly.

In a typical mouse tracking experiment (e.g., Spivey et al., 2005; Dale et al., 2007; Freeman et al., 2008; Farmer et al., 2009; Coco & Duran, 2016), participants are asked to solve a cognitive task and indicate its solution by moving a mouse-controlled cursor from the starting position at the bottom center of the computer screen to the correct response location. There are usually two response options, located on the left and right side in the upper screen region. The trajectory of the mouse cursor is recorded during the response. Typically, the certainty over the task solution is reflected by the movement deviating somewhat into an intermediate direction, that is, toward the incorrect option. The impact of the alternative but incorrect response option on trajectory shape

is used as a measure for the evolution of the cognitive task over time.

For instance, Dale et al. (2007) asked their participants to sort animal names into one of two categories. A trial started with two category labels written in the top corners of the screen, for instance, "mammal" on the right and "fish" on the left. Two seconds later a button appeared at the bottom center of the screen, which participants had to click. Next, an animal name appeared at the same position, and participants had to click on the correct category name for that animal. In control trials, the animal was a typical example for one of the categories (e.g., "dog" for "mammal"). In experimental trials, the animal was an atypical example for one of the categories (e.g., "whale" for "mammal").

Dale et al. computed trajectory divergence as the difference between the horizontal trajectory coordinates in the control and experimental condition at each point in time. It was found that trajectories deviated toward the incorrect alternative more strongly in the experimental condition, starting after approximately half of the total movement time, indicating more intense competition when the animal posed an atypical example for the correct category and shared properties with animals belonging to the alternative option. This was interpreted as evidence that in categorization alternative categories are partially active and compete with the correct one over time.

Mouse tracking was furthermore used to gain insight into many other cognitive domains, such as social categorization (Freeman et al., 2013; Freeman & Ambady, 2011; Freeman et al., 2008; Cloutier et al., 2014), processing of grammatical aspect (Anderson et al., 2013), vowel discrimination (Farmer et al., 2009), cognitive flexibility (Dshemuchadse et al., 2015), intertemporal decision making and delay discounting (Dshemuchadse et al., 2013; Scherbaum et al., 2013, 2016), multitasking (Scherbaum et al., 2015), stimulus-response compatibility (Flumini et al., 2014), lexical decision (Barca & Pezzulo, 2012), and response selection (Wifall et al., 2017). The vast majority of mouse tracking studies employed the standard two-choice paradigm (Hehman et al., 2015). Variants have been used as well (e.g., Cloutier et al., 2014; Wifall et al., 2017; Farmer, Cargill, et al., 2007; Farmer, Anderson, & Spivey, 2007; Anderson et al., 2013; Scherbaum et al., 2013; Koop & Johnson, 2011) but generally stayed in the same methodological frame.

An advantage of mouse tracking and similar techniques over traditional measures such as reaction time or accuracy is that it provides a continuous stream of data, which allows to link behavior to ongoing dynamics of cognitive processes as described above (Freeman et al., 2011). This allows, for

instance, to distinguish whether a cognitive process evolves in discrete steps or continuously over time. This sets mouse tracking apart from eye tracking, which can be used in similar paradigms but is (largely) restricted to the analysis of discrete fixation points (Magnuson, 2005), so that it is not easily possible to infer whether competition is of a graded or discrete nature (Farmer, Anderson, & Spivey, 2007). It has also been demonstrated that trajectory deviation may uncover effects which response times do not reflect (e.g., Koop & Johnson, 2011).

The main dependent variable in mouse tracking is usually the degree of trajectory deviation from the direct (i.e., straight) path between the starting position and the correct response option (other measures have been used as well; see Hehman et al., 2015, for an overview). This deviation can be compared between conditions either for each point in time or as an aggregated scalar index, most commonly the maximal Euclidean distance from the direct path or the area under the curve between trajectory and direct path (Freeman & Ambady, 2010). Competition between response options is judged to be greater in time windows or conditions where deviation is larger. The degree of competition at each point in time is in turn assumed to arise from the momentary state of the cognitive task being solved, so that observed patterns may inform about underlying mechanisms (Freeman et al., 2011).

Another common method is to look for signs of bimodality in the distribution of trajectories over curvature, where curvature is usually measured as one of the scalar indices of deviation mentioned above (Freeman & Dale, 2013; Hehman et al., 2015). This is done to test whether deviation observed in a mean trajectory may in fact have arisen from averaging two (or more) distinct populations of trajectories, one with and one without deviation. This latter state of affairs is usually interpreted to indicate not gradually evolving competition between response options over time, but a discrete decision being made early and later corrected abruptly in some trials (Farmer, Anderson, & Spivey, 2007, obtained such a pattern in a visuomotor-control study, contrasting it with a condition that yielded weaker but sustained deviation). In accord with this, the presence versus absence of bimodality is often used as an indicator for the plausibility of competing theories about the mechanisms behind the cognitive task in question, for instance, to differentiate between stage-based and continuous accounts of word recognition (Spivey et al., 2005) or syntax-first and constraint-based theories of sentence parsing (Farmer, Anderson, & Spivey, 2007).

1.6.1 Novel application in the current work

The experiments described in Chapter 3 aim to measure motor signatures of the embodied neural processes that ground spatial language in the visual world. For this, the well-established method of computer mouse tracking is used. Compared to previous applications, however, the paradigm employed in the current work combines multiple novel or rarely used aspects.

In the experiments, participants read a spatial phrase that describes a relation between two colored items, such as "The green item to the left of the red one.", and then see a visual scene containing twelve colored items, including the described pair. The computer mouse has to be moved from a starting point to the target item (here, the green one) while the trajectory is recorded. The other items include variously colored fillers, one or more distractor items of the same color as the target and, in some experiments, items sharing the color of the reference item (here, the red one).

Consistent with the model in Chapter 2 and the framework of embodiment outlined in the previous section, it is assumed that the processes required to determine which items the spatial phrase refers to operate directly on the grounded substrates that also represent the percepts evoked by the visual scene, guided by the language input. In the course of this, the focus of neural processes must shift between between the locations of these items. The main goal of the experiments is to demonstrate a reflection of these shifts in motor outcome, to provide evidence for the postulated embodied mode of linking language to the visual environment.

One important difference from previous mouse tracking work is that, here, spatially localized effect sources are situated on both sides of the direct path to the target, and that the individual impact of each of these is specifically examined. Usually, mouse tracking studies consider only a single source of potential attraction (mostly the sole alternative response option), so that any deviation can be interpreted in relation to the location of that source (but see Scherbaum et al., 2015).

Here, trajectories are expected to be biased by items of the same color as the target, the reference item and, finally, the screen center as a default response direction (the spatial term is expected to exert an additional influence). Where each of these effect sources is located varies from trial to trial. The various biases are expected to superimpose in each individual trajectory and, due to the variable placement, to do so in a different manner in each trial. Therefore, net trajectory biases can potentially go in either direction, or even change directionality over movement time. To nonetheless disentangle

the impact of potential effect sources, an advantage of mouse tracking is leveraged that is rarely emphasized, namely the ability to differentiate the source of biases based on their directionality.

Note that a consequence of the expected superimposition of multiple effects, and in particular the possibility that the direction of deviation may change within a single trial, is that aggregated measures such as maximum deviation from the direct path or area under the curve are less informative in the current context than in previous approaches. To nonetheless be able to test for bimodality, curvature is assessed using a custom method (see Section 3.1.1) whose outcome is not affected by the distance from the direct path.

A second difference to previous experiments is that not only the sides, but also the locations of potential targets and other effect sources are variable from trial to trial.[1] Moreover, where these will occur in a given trial is not known to the participants before movement onset. Thus, movement planning to new locations must occur fully in parallel with task processing.

Scherbaum et al. (2013; see also Scherbaum et al., 2016) similarly used variable placement but showed response options before movement onset, although time-pressure was induced otherwise. Other studies showed response options only after movement onset, or delayed the presentation of additional stimuli that fully specified the cognitive task until that time, but did not use varying response locations (e.g., Scherbaum & Kieslich, 2017; Dshemuchadse et al., 2013).

Finally, another rarely explored aspect is that response space and task space are strongly overlapping in the current paradigm (but see Farmer, Cargill, et al., 2007, for an experiment in a related spirit). This is because the processes hypothesized to underlie the grounding of spatial language operate within sensorimotor representations of the same space in which response actions are specified. This stands in contrast to previous mouse tracking investigations of higher cognitive tasks, that assigned solutions of abstract cognitive tasks to response locations in an arbitrary manner (e.g., through word labels or learned associations).

In the current paradigm, influences on trajectory shape are expected to arise from the operation of task-relevant processes on sensorimotor representations, and not (only) from arbitrary links between the certainty of competing solutions of an abstract cognitive task and spatial locations. As a consequence,

[1] The location of the correct target is in fact variable between only four positions across trials, but this is successfully masked by the position of all other items being highly variable (as later described, participants reported not to have noticed the fixed target locations; see Section 3.1).

it is expected that not only alternative response options affect trajectories, that is, items of the same color as the target item, but also items only implicated in solving the cognitive task (especially the reference item).

1.7 Evaluating spatial relations

Language enables communication about shared environments, such as pointing out a relevant object to direct the recipient's attention toward it. One way of doing this is to name unique features or object identity. For instance, there is only one object in Figure 1.3 to which "the orange block" may refer. This is not sufficient, however, when multiple similar or identical objects are present. "The green block" may refer to either of two objects in Figure 1.3. In such cases, spatial language can verbally disambiguate which object is meant. Projective relations are commonly used for this, such as those described by "left of", "right of", "above", and "below". For instance, a single green block in Figure 1.3 is uniquely specified by "the green block to the right of the yellow block". Projective relations may be unambiguous even when neither of the involved objects is unique. For instance, the same green block as before is uniquely specified by "the green block below the red block", even though there are two red and two green objects in the scene.

Figure 1.3: Verbally referring to a specific green or red object in this scene is possible only with the help of spatial language.

Relational phrases like this consist of three components: a *target object*, corresponding to the green block in the latter example; the relation itself, denoted by the *spatial term* "below" in the example; and a *reference object*, which corresponds to the red block in the example.[2]

The model and experiments described later examine relations such as the one described in the example above. These are referred to as deictic relations. In deictic relations, the reference frame of the viewer defines which directions correspond to left, right, above, below, and so on (and not the intrinsic reference frame of the ref-

[2]In other parts of the thesis, the reference object and the target object are instead called reference item and target item, respectively, or simply reference and target, for brevity. This is to accommodate the fact that model input and experimental displays are composed of two-dimensional stimuli instead of real-world objects.

erence object, which might be different, e.g., a car has its own front, back, left, and right), while the origin of the reference frame is centered on the reference object (Logan & Sadler, 1996).

Linking a spatial phrase that describes a deictic relation to a configuration of objects in the visual environment requires multiple computational steps, which have been analyzed in a seminal study by Logan & Sadler (1996):

First, the two arguments of a relation, which are initially represented only in a non-perceptual form, must be linked to the locations of the corresponding objects in a perceptual representation. Logan & Sadler (1996) refer to this as spatial indexing. Since the roles of reference and target object are not interchangeable, spatial indexing must be organized such that reference and target location are linked to the correct arguments. Second, the parameters of the reference frame must be set. For deictic relations, this means that the origin of the reference frame is centered on the reference object, while its other parameters, including scale, direction, and orientation, remain congruent with the viewer's reference frame. Third, a spatial template must be imposed on the reference object within the adjusted reference frame. The spatial template is specific to the relation in question and indicates the goodness of fit for different locations in space relative to the reference object. Finally, the goodness of fit must be assessed for the target object by comparing its position to the spatial template. Note that the order in which these steps occur is assumed to vary depending on the task, for instance, judging whether a given relation applies as opposed to selecting a relation to describe the location of an object (Logan & Sadler, 1996).

This framework refutes the subjective intuition that spatial relations are instantly available throughout the visual field, which is already problematic due to the combinatorial explosion of possible relations when many objects are present (Franconeri et al., 2012). In line with this, empirical evidence suggests that much of relation processing involves the sequential processing of objects and relational pairs.

A first hint that this may be required, especially for the spatial indexing step, can be derived from the classical notion that focused attention is required to localize features in the visual environment (Treisman & Gelade, 1980). Event-related potentials (namely, an enlarged N2pc component) similarly suggest that selective attention is engaged more strongly for tasks where the location of a visual target is to be reported compared to a detection task (Hyun et al., 2009).

Franconeri et al. (2012) report data consistent with this for the case of spa-

tial relations. Participants completed a task where they saw two different stimuli (along with two irrelevant fillers) and judged the spatial relation between them. The two stimuli were aligned horizontally, but one was located to the left of visual fixation, and the other to the right of it, so that any attentional shift toward the stimuli would be equivalent to shifting attention into one of the visual hemifields. Event-related potentials were measured during the task, allowing to assess whether attention was shifted into one of the hemifields (through the laterality of the N2pc component). It was found that when the display was shown, attention shifted first toward one, and then toward the other stimulus. This pattern occurred even though participants were instructed to judge the relation by focusing on both items at the same time, and even when another visual task was completed simultaneously. This suggests that either stimulus needs to be sequentially selected to evaluate the relation between them.

In an eye-tracking study by Yuan et al. (2016), participants saw visual displays with two differently colored stimuli that were vertically aligned, such that they could be viewed as instantiating an 'above' or 'below' relation. Presentation time was brief, such that only one or two saccades were possible during presentation. After each display, the participants performed either a spatial recall task, in which they indicated for one of the colored items whether it had been in the upper or lower relative position, or they performed a non-spatial recognition task, in which they indicated which of two colored items had been present in the display. Overall, it was found that the direction of eye movements between items influenced response times in the spatial recall task, while response time in the identification task was influenced by which item was selected in a first fixation. If the spatial recall task queried the item toward which a saccade had occurred starting from the other item (or from a position between the items), then response time tended to be lower compared to the case in which the other item was queried. For the identification task, the firstly fixated item was faster responded to when queried. This suggests that attentional shifts between items, as evidenced by eye movements, facilitate the encoding of spatial relations, and that the order in which items are focused may not be arbitrary.

However, the role of shift order is not fully settled. In another eye-tracking study (Burigo & Knoeferle, 2015), participants listened to a relational phrase and verified it against a visual display containing the two involved objects and an irrelevant competitor. Shifts from the reference object toward the target object were found, but these did not always occur, and even when they

were actively prevented, for instance, by removing the target object before the shift could take place, accuracy was not affected. In this study, however, playback of the relational phrase was started with scene onset, and eye movements toward the visual objects generally followed the order of mention, as in classical visual world studies (e.g., Cooper, 1974; Tanenhaus et al., 1995). Due to the sentence structure used, for instance, "The box is above the sausage", this meant that the target object and the reference object had already been focused in the converse order before any shift from reference to target would occur. This seemed to suffice in most cases to judge the relation, since overall accuracy was high.

As noted by the authors of the above study, this overall pattern is only partly compatible with the shift direction assumed by a well-known model of the low-level neural mechanism that could underlie computing spatial relations (essentially a population vector describing the attentional shift from reference to target object; Regier & Carlson, 2001). In line with this, a reversed version of that low-level model where the shift instead occurs from target to reference object has been shown to work equally well (Kluth et al., 2016). Thus, together, the evidence suggests that attentional shifts are implicated in relational judgments, but it is not fully clear whether order plays a decisive role (and covert attention shifts can, of course, not be ruled out in eye-tracking studies).

Another aspect that appears to invoke sequential processing is the presence of multiple candidate pairs. In experiments by Logan (1994; see also, Moore et al., 2001; for review, see Carlson & Logan, 2005) participants saw visual displays with multiple item pairs and reported the presence or absence of a target pair that was defined by a relational phrase (e.g., by "dash above plus"). If present, the target pair was placed among distractor pairs that instantiated the opposite relation (e.g., dashes below pluses). Search time rose steeply with the number of distractor pairs (by approximately 85 ms per item when the target was present). Search time slopes were flat, in contrast, when distractor pairs consisted of all dashes or all pluses, which Logan (1994) attributed to attentional pop-out of the discrepant item in the target pair. Interestingly, however, the pop-out did not appear to help processing the relation of the pair containing the discrepant item: Deciding whether the appropriate relation was present in the display still took more time than only deciding whether a discrepant item was present in the display, which was probed in another condition. Together, these results suggest that attentional allocation is required but not sufficient to process relational pairs, which instead seems

to involve additional steps (Logan, 1994).

A final relevant point for the current work is the shape of spatial templates, which has been assessed through acceptability ratings (asking participants to rate how well the position of a target object is described by a given spatial term) and production tasks (having participants mark positions that instantiate a given relation or letting them describe a scene). For instance, Hayward & Tarr (1995) asked participants to describe visual scenes in terms of spatial relations that contained a reference object and a target object. It was found that vertical terms, such as "above", were used most often when the target object was located along a vertical axis extending from the reference object into the respective direction, while the use of these terms declined with rising angular distance from that axis. A similar picture emerged for horizontal terms, like "left of", and in an applicability rating task. Logan & Sadler (1996) obtained similar results when participants rated applicability or placed target objects to instantiate given relations.

1.8 Dynamic Field Theory

Dynamic Field Theory (DFT; Schöner, 2008; Schöner et al., 2015) is a theoretical framework built on the notion that activation patterns within neural populations are directly linked to macroscopic events in perception and behavior (Erlhagen et al., 1999; Jancke et al., 1999; Bastian et al., 2003). It captures neural activation patterns at the population level in the form of Dynamic Fields (DF; the term 'field' will be used interchangeably), and allows to simulate the evolution of these patterns through numerical simulation. The way in which DFT describes neural activity is rooted in neurophysiology through evidence on how the nervous system represents attributes of sensory stimuli and motor actions. Based on this and extrapolating to more abstract domains, DFT captures not only elementary perceptual decisions but proposes a dynamical systems perspective on higher cognitive capabilities, providing the tools for an operational process account of cognitive processes.

1.8.1 Roots in neurophysiology[3]

It has been shown that individual neurons in the nervous system respond to very specific aspects of behavior, perception, or cognition. A classical example from the visual cortex are neurons that respond differentially depending on the orientation of a line within some specific region on the retina (Hubel & Wiesel, 1968). The concept of neural tuning refers to the fact that such neurons are active whenever a given sensorimotor or cognitive parameter value is within a specific range, and that the magnitude of activation depends on the distance of the current value from a preferred one which drives the cell most strongly. A neuron's tuning curve makes this concrete, by describing the neuron's response for each value along a particular parameter dimension.

Neural tuning curves are often well-approximated by Gaussian curves or similar functions, which peak around a preferred value (sometimes multiple ones) and fall off to either side of that value. Examples for this have been abundantly reported in the classical and recent literature, for instance, in the form of tuning to the position of stimuli on sensory surfaces, such as the location of a visual stimulus on the retina or a tactile stimulus on the skin, in which case tuning curves are equivalent to receptive field profiles (Jones & Palmer, 1987; Sherrington, 1906), tuning to motor space, such as the target position of a saccade (Lee et al., 1988) or the direction of a hand movement (Georgopoulos et al., 1982), tuning to visual feature dimensions like orientation (Hubel & Wiesel, 1968) or color (Conway & Tsao, 2009), or higher-level examples such as tuning to boundary curvature at specific angular positions relative to an object center (Pasupathy & Connor, 2001), to numerosity of visual objects (Nieder & Miller, 2003), or even with respect to conceptual spaces (Gotts et al., 2011).

For any dimension coded in this way there usually exist whole populations of neurons with diverse preferred values. Together with the graceful decay of tuning curves on either side of the preferred value this entails that the curves of different neurons overlap, so that an ensemble of neurons becomes active for any specific parameter value, say a particular reaching direction (Georgopoulos et al., 1988), or a saccade end point in retinal space (Lee et al., 1988). This gives rise to the population coding hypothesis (Erickson, 1974; Georgopoulos et al., 1983), which holds that information about a cur-

[3]This section includes modified material previously published in *Neural Fields*, Coombes, S., beim Graben, P., Potthast, R., & Wright, J. (Eds.), A Neural Approach to Cognition Based on Dynamic Field Theory, 2014, pp.319-339, Lins, J., & Schöner, G. (© Springer-Verlag Berlin Heidelberg 2014). With permission of Springer.

rently coded parameter is indeed represented jointly by all active neurons, with each neuron 'voting' for its preferred value to the degree of its level of activation.

In support of this hypothesis it has been demonstrated (Groh et al., 1997; Georgopoulos et al., 1988; Lee et al., 1988) that an average over the preferred values of neurons activated by a given parameter value accurately represents the coded value when first weighting each preferred value by the respective neuron's degree of activation. Importantly, these studies also show that even weakly activated neurons with preferred values very different from the specified value impact the perceptual or motor outcome.

Subsequent work has further qualified how population activation represents sensorimotor parameters, by demonstrating that the shape of the distribution is meaningful beyond the mere average of preferred values. This type of information is preserved by the distribution of population activation (DPA), whose basic rationale is to compute activation distributions by summing entire tuning curves instead of preferred values, weighting each curve by the degree of activation the respective neuron exhibits in response to a probed parameter value (Erlhagen et al., 1999; Bastian et al., 2003; Jancke et al., 1999; Cisek & Kalaska, 2005). The result is an activation distribution over the same parameter space that was sampled experimentally to obtain the tuning curves. The shape of the DPA correlates, for instance, with the certainty over reaching targets and the associated latency of a motor response (Bastian et al., 2003), or with the concurrent representation of multiple possible targets through multimodal distributions (Cisek & Kalaska, 2005). This provides the basis for the stance underlying DFT that distributions of activation are the appropriate level of consideration to capture macroscopically relevant behavioral, sensory, and cognitive decisions.

DFT also builds on neurophysiological findings that show lateral interactions within populations of neurons that are sensitive to the the same sensorimotor dimensions, meaning that patterns of activation in such populations are shaped not by feed-forward input alone (Schneegans, Lins, & Schöner, 2015). The interaction of cells typically depends on their distance in the coded feature space, with lateral connections tending to be excitatory between neurons with similar preferred values and inhibitory between neurons coding for very different values, as has been demonstrated, for instance, in the cat visual cortex (Ts'o et al., 1986), and in the monkey motor cortex (Georgopoulos et al., 1993). Effects of lateral interactions on DPAs have been observed in the cat visual cortex, in the form of weakened activation peaks for the case of two

simultaneously presented spatially remote stimuli (long-range inhibition) and temporarily enhanced activation peaks for closely spaced stimuli (short-range excitation; Jancke et al., 1999).

To summarize, the shape of the DPAs carries information that observably impacts behavior, with activation peaks pertaining to macroscopically relevant perceptual or behavioral conditions. This provides the foundation for the stance of DFT that distributions of activation in neural populations are an appropriate level for describing and simulating how neural activation represents perceptual and motor parameter spaces and, extrapolating from the immediate evidence, cognitive dimensions more remote from the sensorimotor surfaces. The particular regimes of lateral interaction used in dynamic neural fields (DFs), on the other hand, are paralleled by physiological findings about horizontal connectivity in neural maps.

1.8.2 Dynamic Neural Fields[4]

DFT describes the evolution in time of activation patterns in neural populations, in a manner linked to neurophysiology through the DPA method. Activation patterns are modeled as DFs that are defined over continuous metric dimensions and evolve continuously in time.

Special focus is laid on modeling lateral interactions within the fields that endow them with a particular set of stable attractor states. These stable states correspond to meaningful representational conditions, such as the presence or absence of a particular value along the coded dimension. Instabilities that lead to switches between the different stable states are brought about by sufficient changes in the configuration of the external input a field receives from sensory surfaces or other neural substrates. Different DFs may vary with respect to the exact configuration of interaction parameters as long as the stability properties characteristic for DFs are retained. The differences in dynamic behavior that result from different sets of interaction parameters are decisive for each field's specific functionality within the context of a larger neural architecture.

The particular mathematical form of field dynamics adopted by DFT has first been analyzed by Amari (Amari, 1977; see also Grossberg, 1978; H. R. Wil-

[4]This section includes modified material previously published in *Neural Fields*, Coombes, S., beim Graben, P., Potthast, R., & Wright, J. (Eds.), A Neural Approach to Cognition Based on Dynamic Field Theory, 2014, pp.319-339, Lins, J., & Schöner, G. (© Springer-Verlag Berlin Heidelberg 2014). With permission of Springer.

son & Cowan, 1972):

$$\tau \dot{u}(x,t) = -u(x,t) + s(x,t) + h + \int k(x-x')g(u(x',t))\,\mathrm{d}x' \qquad (1.1)$$

Here, $u(x,t)$ is the field of activation, defined over the metric dimension, x, and time, t. From a neurophysiological viewpoint, the activation, u, can be interpreted as a correlate to the mean membrane potential of a group of neurons. The time scale of the relaxation process is determined by τ. The field has a constant resting level, h, and may receive localized patterns of external input, $s(x,t)$. The last term describes lateral interactions between different field sites. Here, g is a sigmoid function implementing a soft threshold for field output, and k is an interaction kernel that specifies the strength of interactions between different field sites as a function of their metric distance. The kernel typically has a Mexican hat shape, implementing local excitation and surround inhibition, usually with added global inhibition. This means that field sites coding for similar parameter values excite each other, while mutual inhibition predominates between field sites that code for very different values. The sigmoidal threshold function ensures that only sufficiently activated field sites generate output and impact on other sites or downstream substrates (see below). The field output can be viewed as corresponding to the mean spike rate of a group of neurons.

In the absence of supra-threshold activation, no output is generated. In this case, the entire field relaxes to the stable attractor that is set by the resting level (which usually resides well below the output threshold). A flat distribution indicates the absence of any specific information about the coded dimension.

When weak, localized input is applied, the attractor at the respective field site is shifted toward the output threshold. As long as the threshold is not reached, though, the field state remains purely input-driven and activation thus simply traces the shape of the input (Figure 1.4a). Although there is now some structure to the distribution, this state still indicates the absence of conclusive information.

If, in contrast, the localized input is sufficiently strong to push a section of the field above threshold (or near it, see below), output is generated and lateral interaction kicks in. For the usual DF, parameters that define the interaction kernel reside within a range which ensures that lateral interaction promotes the formation of a localized peak of activation (Figure 1.4b). Namely, local excitation further elevates activation around the input position, whereas

more distant field sites are depressed by global inhibition and/or surround inhibition, which prevents the peak from dispersing. Due to these properties the peak is referred to as *self-stabilized*.

The transition from a sub-threshold solution to a self-stabilized peak is called the *detection instability*, since it corresponds to the decision that a coherent, well-defined item is present in the input stream. Peaks are units of

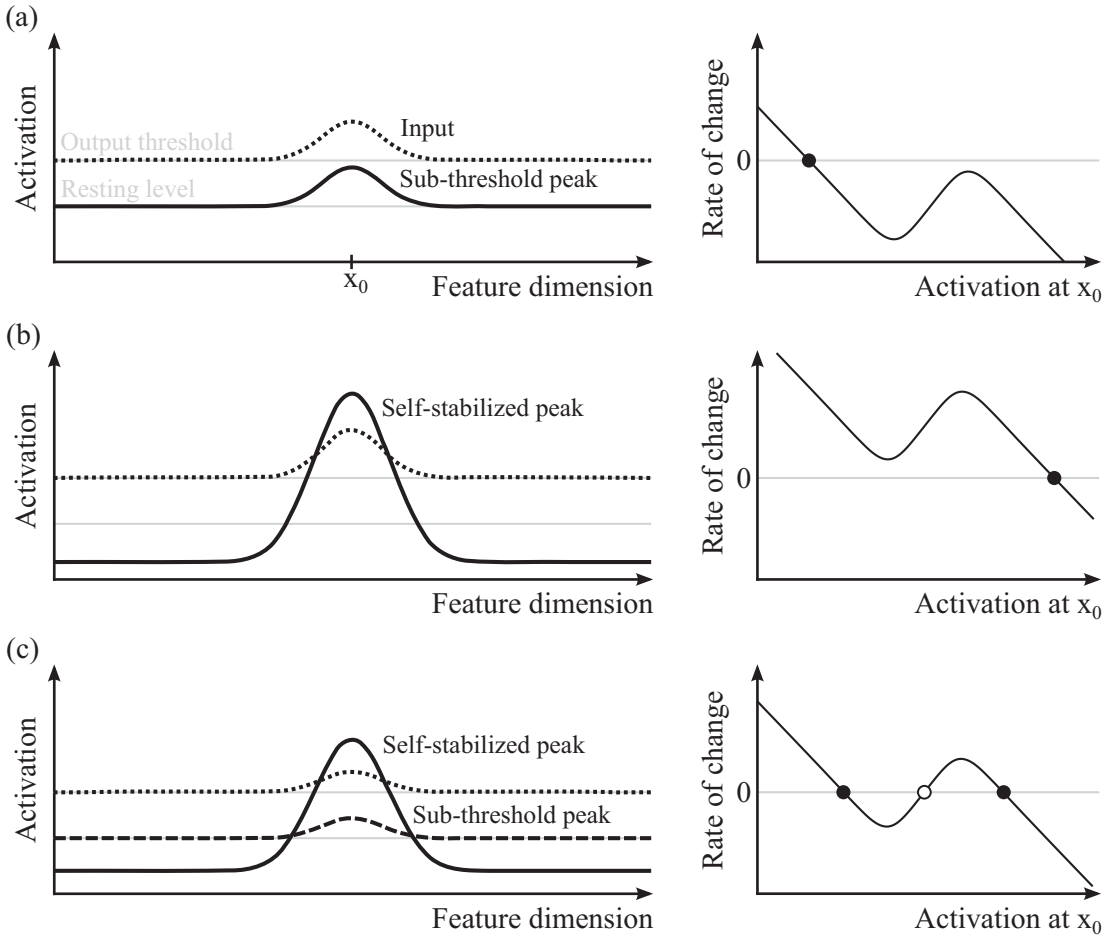


Figure 1.4: Left column: Stable states reached by dynamic neural fields (solid lines and long dashed line) as a result of localized Gaussian inputs of different strengths (dotted lines). Right column: Corresponding plots of the rate of change as a function of activation at the peak position, x_0 (note that these plots are only approximate, as they do not take into account the impact of other field sites on the rate of change at x_0 via lateral interactions). Attractors are marked by filled dots, repellors by open dots. (a) Weak input results in a purely input-driven sub-threshold peak, which is a monostable attractor state. (b) High levels of input that bring activation above threshold result in output generation and lateral interactions, thus leading to a self-stabilized peak. This state as well is monostable. (c) For intermediate input strengths the system reaches a bistable state. The current state then depends on the system's prior state. Here, the self-stabilized peak (solid line) corresponds to the attractor on the right side, which is reached from high levels of activation. The sub-threshold peak (long dashed line) corresponds to the left attractor, which is reached from low levels of activation. Reproduced from *Neural Fields*, Coombes, S., beim Graben, P., Potthast, R., & Wright, J. (Eds.), A Neural Approach to Cognition Based on Dynamic Field Theory, 2014, p.421, Lins, J., & Schöner, G. (© Springer-Verlag Berlin Heidelberg 2014). With permission of Springer.

representation in this sense, indicating that a particular parameter value is present in the sensory environment, as part of a motor plan, as the contents of memory, or in the context of another cognitive process. The encoded value itself – what is being perceived, planned, or memorized – is specified by the position of the peak along the metric dimension.

The detection instability prevents decisions from fluctuating, by transforming even near-threshold activation into macroscopic peaks (due to the sigmoidal threshold function local excitation becomes effective even when activation is somewhat below threshold). This enables, for instance, attending a weak visual stimulus over some time even in the face of random fluctuations or perturbations of activation, which may arise in the nervous system due to the inherent variability in neural firing or as the result of currently ongoing but unrelated neural processes. Such magnification of microscopic decisions does not occur in purely input-driven systems, where a near-threshold input would allow fluctuations to push activation below and above threshold from one moment to the next. The importance of amplifying sensorimotor decisions to macroscopic levels has been concretely demonstrated in the context of perseverative reaching in young infants, through a DF model that captures data collected in the A-not-B paradigm of developmental psychology (Dineva & Schöner, 2018). The dynamic instabilities described in the following similarly serve to amplify, suppress, and sustain neural decisions and thus enable cognition to decouple from the immediate sensory environment to a degree, supporting stability in noisy environments laden with equally salient options.

It is called the *reverse detection instability* when a self-stabilized peak vanishes. This happens when the localized input that brought about the peak is sufficiently reduced in strength. For example, when the input is removed entirely, the peak attractor becomes unstable and disappears, while the resting level attractor reappears, to which the system then relaxes. Decreasing the input strength successively will also eventually trigger the reverse detection instability, but local excitation to a degree shields existing peaks from decaying. The system will thus stick to the detection decision across a range of input strengths that would not have triggered the detection instability in the first place. The system is bistable over this range, with the peak attractor and the input-driven attractor coexisting (Figure 1.4c). The field state then depends on which basin of attraction it resided in prior to the change of input strength.

This dynamic behavior is known as hysteresis and, similar to what has been described above for the detection instability, stabilizes decisions against

fluctuations. In line with this, signatures of hysteresis are a common finding in behavioral experiments (e.g., in apparent motion perception, Hock et al., 1997; for review, see Hock & Schöner, 2010).

Besides self-stabilized peaks, the other fundamental attractor state in DFT is the *self-sustained* peak. These peaks occur instead of self-stabilized ones when (net) local excitation in a DF is so strong that it can by itself prevent the supra-threshold attractor from vanishing once a peak has been established. In this regime, peaks decay only when the level of activation is sufficiently decreased, locally or globally, by external inhibitory input or by endogenous inhibitory interactions. Otherwise, self-sustained peaks may persist indefinitely in the absence of input. The self-sustained regime enables DFs to support the functionality of the neural process of working memory (Johnson, Spencer, Luck, & Schöner, 2009; see also Fuster & Alexander, 1971).

Further types of instabilities may occur when multiple localized inputs impinge on a field simultaneously — as is the rule in natural environments that are richly structured, cluttered with stimuli, and offer a variety of behavioral goals and movement targets that compete for processing and behavioral impact.

If two localized inputs are so widely spaced that they interact only via global inhibition, selection will occur, provided the strength of global inhibition is large. Note that this also requires that the two inputs differ in strength at least somewhat or that such an imbalance is introduced through noise (which is implemented in DFs as Gaussian white noise; see Schöner, Reimann, & Lins, 2015). Since more self-excitation will occur around the location that is activated more strongly, the height of the corresponding peak is further elevated above that of its competitor. The ensuing increase in global inhibition suppresses the weaker peak, eventually reducing it to an input-driven bump. This is referred to as the *selection instability*. The single-peak state resulting from selection is bistable, with both peak attractors coexisting (Figure 1.5a).

Depending on the metrics of the inputs, however, multiple inputs may also lead to fusion, analogous to behavioral data about averaging saccades to the midpoint between closely spaced stimuli (Ottes et al., 1985; this has in fact been captured by a DF model of the superior colliculus; Wilimzig et al., 2006). Two inputs result in a single peak at an average position if they are so close to each other that the regions of input-induced activation are subject to mutual lateral excitation. The activation then propagates from the input positions towards the center between them, eventually forming a single peak (Figure 1.5b). The fused peak state is monostable for very close inputs, but

becomes bistable when inputs are moved apart. If the distance is increased even more, the attractor of the fused state becomes unstable and disappears, and the field relaxes to a selection state as described above (or to a two-peak state if the field regime is not selective; in this case, surround inhibition may lead to repulsion between peaks, which has been used to explain drift in spatial working memory; Johnson & Simmering, 2015).

While the number of peaks supported by a field operating in a selective regime is limited through inhibition, even such fields are not necessarily constrained to a single peak. The maximum number of peaks that a DF can support depends on the balance of excitation and inhibition, with selection becoming more likely as the sum of inhibition generated by the existing peaks becomes larger. This is particularly relevant for modeling explicit capacity limits in cognition, such as those in working memory or attentional function.

This section has so far dealt only with one-dimensional fields, but DFs may also be defined over multiple dimensions. Each dimension then corre-

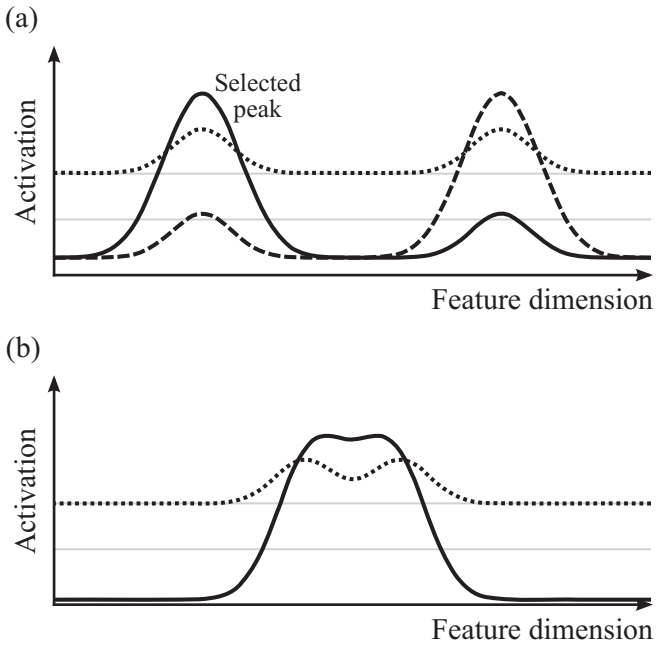


Figure 1.5: Stable states reached by dynamic neural fields (solid lines and long dashed line) as a result of different patterns of localized Gaussian input (dotted lines). (a) Competition between peaks occurs when two inputs are applied at distant positions. Only at one location is a self-stabilized peak formed (solid line), while the other is suppressed by inhibition. The state resulting from this selection decision is bistable, with the alternative state (long dashed line) continuing to coexist as an attractor. Which state is reached depends on the field's prior activation history, imbalances between the inputs, and noise. (b) Two close inputs can result in a monostable fused peak state, with a single peak at an average location between the inputs. Reproduced from *Neural Fields*, Coombes, S., beim Graben, P., Potthast, R., & Wright, J. (Eds.), A Neural Approach to Cognition Based on Dynamic Field Theory, 2014, p.424, Lins, J., & Schöner, G. (© Springer-Verlag Berlin Heidelberg 2014). With permission of Springer.

sponds to another sensory, motor, or cognitive feature. For instance, for a hypothetical population of neurons responsive to changes along only a single feature, say color hue, the distribution of activation can be captured by a one-dimensional field, while a population jointly responsive to both color hue and position in retinal space requires a three-dimensional field (color hue and two dimensions of space). A special case is posed by dynamic nodes, which are equivalent to zero dimensional fields, consisting of a scalar activation variable, with recurrent interaction affecting only that same variable.

Very high-dimensional fields are generally avoided, both due to combinatorial reasons (i.e., an explosion of the required number of neurons) and due to the fact that cortical systems likewise tend to be composed of multiple representations each coding for a relatively small number of feature dimensions (Schneegans, Lins, & Spencer, 2015). Note, however, that it is not the aim of DFT to establish one-to-one correspondences between DFs and specific neuron populations in the nervous system. Rather, feature spaces represented in DFT are abstract, and the corresponding neural maps may in reality be interlaced, convoluted, or scattered in cortex. DFT disregards this anatomical layout, focusing on the functional properties of the representations that arise from synaptic connectivity within and between neural populations (Schneegans, Lins, & Schöner, 2015).

Finally, as hinted initially, DFs may be linked to each other to form architectures which perform cognitive tasks that transcend the elementary decisions occurring in each field. Such connections pose an additional source of excitatory or inhibitory input to DFs, besides lateral interactions and sensory surfaces. The impact of synaptic connections from another DF is mathematically captured by a term analogous to the last one in equation 1.1, using similar Gaussian or Mexican hat kernels to mimic synaptic spread.

The connectivity between DFs is organized such that corresponding field sites along a shared feature dimension are linked. This means that a field site representing a given feature value in the source field provides input to all sites that represent the same value in the target field (and to neighboring field sites with decreasing strength, depending on kernel width). If the linked DFs differ in the number of dimensions this connection scheme amounts to collapsing or expanding field output before it is fed into the target field. If the target field is of lower dimensionality than the source, the output is collapsed (i.e., integrated) along the dimensions not represented in the target field. If the target field is of higher dimensionality, the output is taken to be homogeneous along the dimensions not shared by the two fields. A common situation in DF

architectures is that a peak in a lower dimensional field generates output that is projected into a higher dimensional one. The above projection scheme then results in 'ridges' of activation, when a one-dimensional field projects into a two-dimensional one, 'slices' of activation, for projections from one- into three-dimensional fields, or cylindrical columns of activation, when a two-dimensional field projects into a three-dimensional one. The weight matrices of projections from zero-dimensional dynamic nodes onto fields may take various shapes, either homogeneous ones that result in activation in an entire field being shifted up or down, or heterogeneous ones that implement more specific neural operators.

Chapter 2

A Dynamic Field Model of Spatial Language Grounding[1]

2.1 Introduction

The model described in this chapter represents a hypothesis of how a neural system may ground spatial phrases in visual scenes. Spatial phrases are of the form "The green item to the left of the red item"; corresponding visual scenes are shown in Figure 2.1. Grounding is here understood as finding and neurally representing all items in the scene that are contained in the phrase, based on the featural and relational cues the phrase supplies.

The model is based on the embodiment stance outlined in Section 1.5 and the formal framework of DFT outlined in Section 1.8. Thus, the fundamental underlying assumption is that representations of visual items are embedded in continuous, metric sensorimotor representations of the visual world. These are assumed to be contrasted by the amodal, discrete nature of information conveyed in language. The model focuses on how neural substrates supporting these two representational formats may be connected within a neural framework such that ordered processes arise that establish a coherent mapping between language and perception.

The model implements all processes in a pervasively neural manner, by

[1]Material in this chapter is a revised and extended version of M. Richter, Lins, Schneegans, Sandamirskaya, & Schöner (2014). This publication and the model therein, which is described here, represent the outcome of joint work involving the author of the current work (JL), Mathis Richter (MR), Sebastian Schneegans, Yulia Sandamirskaya, and Gregor Schöner. JL and MR developed the model based on collaborative conceptual work and literature survey, with JL focusing on psychological and neural constraints. The technical implementation of the model was done by MR. All authors participated in writing the paper. The simulation results shown in the current section were generated for the current work by JL but are analogous to those presented in the earlier publication.

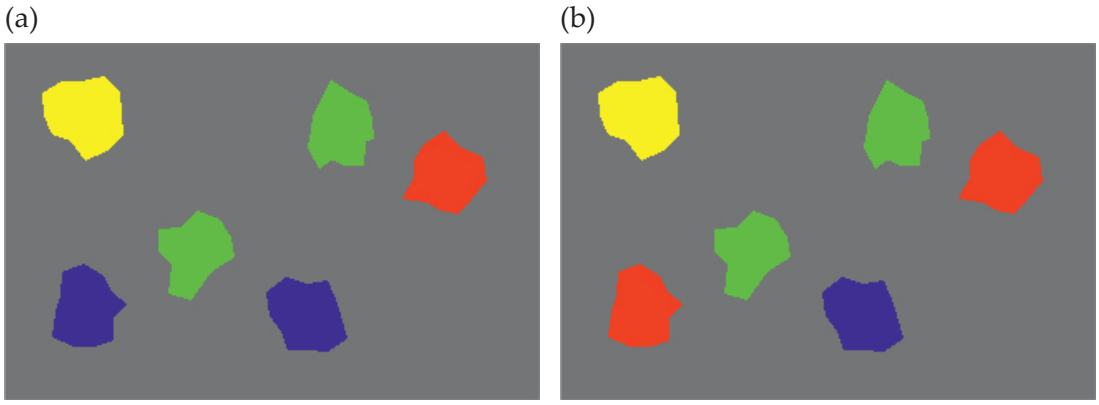


Figure 2.1: These scenes were used as visual input for the model in the simulations presented here. The scenes were supplied to the model together with a pattern of node activations corresponding to the spatial phrase "The green item to the left of the red one".

using only concepts from the theoretical framework of DFT. Phrase representations are realized as discrete neural nodes that interact with continuous dynamic neural fields in which modal patterns of activation are instantiated and selected. The model architecture is one seamless dynamical system that, once supplied with discrete quasi-linguistic and metric visual input, autonomously realizes the processing steps required to ground a spatial phrase.

A previous account realized some of the core mechanisms for evaluating spatial language in accordance with neural principles (Lipinski et al., 2012). The model in this earlier work included the neural substrates required to store reference and target objects, center the reference frame on the location of the reference object through a coordinate transformation, and apply a spatial template to the relational space. Crucially, however, the sequential order of the different operations was controlled through signals supplied from outside the system. Instead of guiding processes in the perceptual substrates based on a neurally represented instantiation of the spatial phrase, processes like selecting an object of a particular color, storing it in the appropriate memory, and matching it to a spatial template, were initiated and terminated by externally supplied and timed sequences of input to the different substrates.

Thus, the model of Lipinski et al. (2012) established the core substrates for performing operations akin to those postulated by Logan & Sadler (1996). It did not include structured representations of linguistic phrases or specify how these may interact with modal substrates to create activation patterns consistent with both of these input sources.

The model described here builds on this earlier work by employing the same types of neural mechanisms for storing the relevant item positions,

transforming coordinate frames, and matching spatial templates. It embeds these within an overarching system of neural control structures that are partly discrete, linking to language and allowing sequential control, and partly modal, allowing to represent the conditions under which a perceptual goal is satisfied or missed. The resulting interplay across substrates organizes the initiation, termination, and resetting of appropriate processing steps, such that spatial phrases can be grounded autonomously, even when multiple candidate items are present in the visual scene. This includes the capability to sequentially test different hypotheses about the possible referents of a spatial phrase when multiple candidates exist (similar to experiments that require searching for visual relations in a cluttered display; Logan, 1994; see *Evaluating spatial relations*, p.22).

Generating ordered processing steps that occur as required by situational sensorimotor constraints is a challenge for neural systems in which activation evolves continuously in time, even though such steps may seem natural in terms of information processing and classical cognitive psychology (as, e.g., in the framework of Logan & Sadler, 1996). The principles according to which the neural control is organized in the model are inspired by earlier work based on DFT that proposed a framework for the autonomous generation of behavioral sequences (Sandamirskaya & Schöner, 2010; M. Richter et al., 2012). The core idea of this work is that elementary processing steps are characterized by aspects that can be implemented in a neural system.

This includes a neurally represented *intention*, which drives the neural structures that are implicated in the processes required to satisfy it. For instance, selecting an object of a particular color requires to instantiate the sensory pattern which signifies that color, in order to drive feature attention. A neural *condition of satisfaction* (CoS) detects the changes resulting from satisfying the intention, indicating the successful completion of a step. Keeping with the example, focusing spatial attention on a particular object in the visual field, guided by the intention, leads to a sensory-driven activation pattern in a perceptual substrate. The neural substrate of the CoS combines the intentional and perceptual patterns to assess their match. Alternatively, a neural *condition of dissatisfaction* (CoD) indicates that a process result is incongruent with the intention. For instance, focusing spatial attention on an item colored differently than what the intention indicates, say, due to another highly salient item being focused, triggers the condition of dissatisfaction. This is achieved through the intention inhibiting the intended values in the substrate of the neural CoD, so that only erroneous selections can trigger it. How these

aspects work together will be made more concrete in the course of the chapter.

This framework is used to enable flexible control of the sequential chain of processes required to solve different tasks in which language and visual spatial relations must be linked.[2]

A guideline in devising the model architecture were the neural constraints imposed by the framework of DFT. Keeping the demand for neural resources within realistic bounds is linked to this, as very high-dimensional fields and other biologically unrealistic concepts are avoided (Schneegans, Lins, & Spencer, 2015). Further constraints arose from the demands of autonomous process organization and from the experimental literature described in Section 1.7. This literature suggests that the selection of reference and target object occurs in sequence, as well as the scrutiny of different relational pairs. That constraint is consistent with the general take of DFT on visual perception, in which visual feature binding, as well as extracting the location of features, is based on spatial attention and subject to fundamental capacity limitations (Schneegans, 2016; Schneegans, Spencer, & Schöner, 2015). This in turn accords to theories and findings from the domain of visual perception that similarly frame spatial attention as a capacity-limited mechanism (Treisman & Gelade, 1980; Hyun et al., 2009; Wolfe, 1994).

It should be noted that fields in the following descriptions are in part labeled in a way that seems to contradict the contention that drawing a border between sensory, motor, and cognitive substrates is problematic in an embodied neural system (see *Embodiment in DFT*, p.10). These labels were chosen to ease description and are based on which function the substrate primarily serves in the architecture at hand, without excluding it being involved in other domains. Similarly, the term 'concept' is used in various places below. It is here employed loosely, without committing to a particular psychological or philosophical definition and without assigning it to any particular physical entity or event; rather it refers to the set of activation patterns and processes required to ground what a given word refers to (see Barsalou et al., 2003, for a similar view on concepts).

[2]Only a restricted set of the model's capabilities is covered here. The main focus will be on those model components that are relevant for the task of grounding a complete spatial phrase in visual input. The model can, without modification, serve other tasks, such as answering questions like "What is to the right of the green object?" by reporting the color of a relational target (M. Richter, Lins, Schneegans, & Schöner, 2014).

2.2 Architecture

The architecture shown in Figure 2.2 is composed of DFs and dynamic nodes that are connected by synaptic projections as described in Section 1.8.2, thus constituting one seamless dynamical system.

The right half of Figure 2.2 shows neural representations of continuous perceptual spaces in the form of one-, two-, and three-dimensional DFs (represented by line plots and color-coded activation patterns). The fields provide the substrate for representing perceptual patterns caused by sensory input and defined over spaces that retain a continuous, metric structure similar to that of the sensory surface (here image space).

Activation patterns in the fields may also be affected by input from the system of dynamic nodes on the left side of Figure 2.2. In the figure, nodes are shown as circles and represent the discrete, amodal format of information that language conveys, such as color terms and spatial relational terms. Each of these nodes has connectivity into the DFs that enables it to evoke an activation pattern there that poses a continuous instantiation of what the node stands for.

The subset of nodes at the very top serves a slightly different role, causing the dynamical system to progress through a given task in elementary processing steps. They control the flow of neural activation in the architecture through connectivity to both the fields and the nodes mentioned above, activating concepts and modulating DF activation levels to enable instabilities in a timely manner as required to successfully complete a given processing sequence.

Input to the architecture comes in two forms. The system of DFs in the right part of Figure 2.2 receives visual input from an image file or camera image. Quasi-linguistic, that is, discrete-type input about the spatial phrase that is to be grounded in the visual scene is supplied to the node system on the left side of the figure. Both types of input are then linked through the ensuing autonomous evolution of activation in the architecture.

The following descriptions move through the different components of the architecture, starting in the top right of Figure 2.2 and proceeding roughly clockwise.

2.2.1 Perceptual and attentional system

Visual input to the architecture takes the form of a distribution of salient colors over image space that is extracted from the input image in a pre-

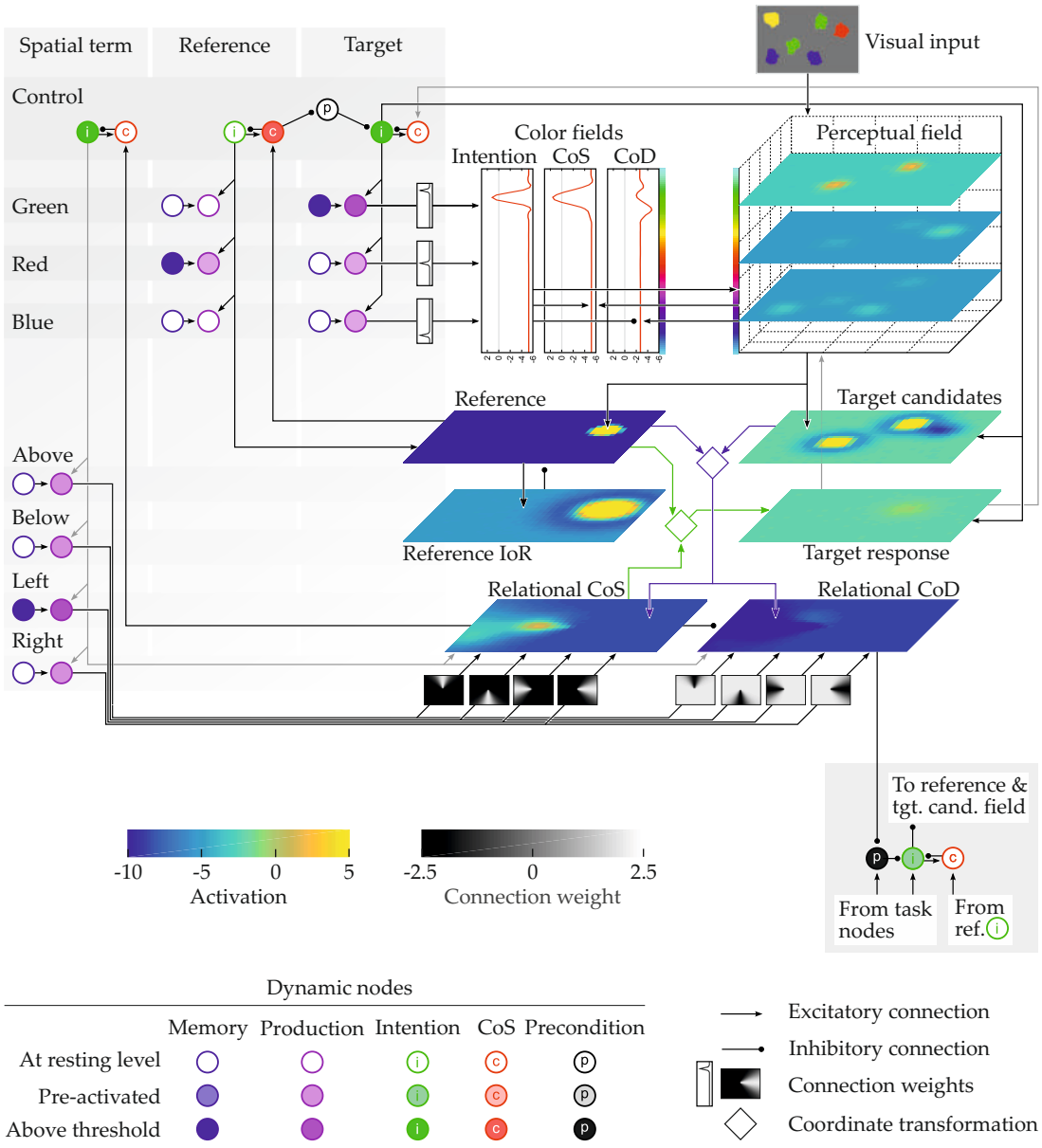


Figure 2.2: Overview of the architecture of the model for spatial language grounding. The figure shows a snapshot of the architecture's activation state as it grounds "the green item to the left of the red item" in the visual scene shown at the top right and in Figure 2.1a. Two- and three-dimensional dynamic fields are shown as color-coded activation patterns, one-dimensional fields as red line plots. For the three-dimensional activation pattern of the perceptual field (top right), slices are shown at color hues green, red, and blue. Dynamic nodes are shown as circles with activation levels indicated by the intensity of the filling color (see table at the bottom). Lines with arrow heads indicate excitatory synaptic connections while inhibitory connections are denoted by lines ending in circles. Note that some connections have been omitted for simplicity, such as those from reference production nodes to the color intention field. Connection weights are coded either by black and white patterns above connection arrows, or as black line plots. Weight patterns of connections are uniform (or scalar) if not indicated otherwise. See text for further details.

processing step outside the scope of the neural model. The input is supplied to the three-dimensional *perceptual field* (top right in Figure 2.2), which is defined over the two spatial dimensions of the image and one color dimension (note that the perceptual field is continuous along all three dimensions despite being illustrated by slices of activation at the color values green, red, and blue). When there are colored objects present in the image, such as in the one displayed in the top right of Figure 2.2, the input produces sub-threshold bumps of activation at the corresponding field coordinates, as is the case in the lower two slices of the perceptual field in Figure 2.2.

Activation in the perceptual field only reaches the output threshold if additional input arrives from the one-dimensional *color intention field* (top middle in Figure 2.2), which is defined over color hue. An activation peak in the color intention field reflects the color of a task-relevant item that is currently of interest, such as the target item, and results in a slice of activation in the perceptual field that is localized along the color dimension. If the activation slice overlaps with an activation bump caused by a visual item in the input image, a peak forms in the perceptual field. If multiple items in the visual input match the color that is represented in the color intention field, multiple peaks may arise in the perceptual field, as is the case in the top slice of the perceptual field in Figure 2.2 with respect to the green items. Bringing peaks above threshold in this manner amounts to bringing objects with a certain feature into the attentional foreground through feature attention.

As Figure 2.2 shows, there are two more one-dimensional color fields that are coupled to the color intention field. These two fields play a role in structuring the processes in other tasks than grounding a spatial phrase for which all three components have been supplied to the node system. They are therefore not relevant in the current context but are briefly explained for the sake of completeness. The *color condition-of-satisfaction (CoS) field* forms a peak when an item of the color currently represented in the color intention field has been brought above threshold in the perceptual field, signalling success in finding an item with the feature specified in the intention field. This functionality is based on converging excitatory input from the perceptual and the color intention field. If an item of any non-matching color comes above threshold in the perceptual field, the *color condition-of-dissatisfaction (CoD) field* builds a peak, signalling that the currently selected item does not match the desired feature and indicating that an additional search pass is required. This is based on inhibition from the color intention field, excitatory input from the perceptual field, and an elevated resting level which allows the latter input to produce a

peak by itself.

2.2.2 Relational system

All remaining DFs in the architecture are two-dimensional spatial representations invariant against color. The *reference field* and the *target candidates field* receive purely spatial input from the perceptual field and serve to represent the location of visual items that are reference or target candidates, respectively, to which the current spatial phrase may refer. The reference field operates in a selective and self-sustained regime, so that only a single reference location is selected from the input supplied by the perceptual field and is stored as a self-sustained peak. The target candidates field is self-sustained as well but is non-selective, thus being able to sustain one or more potential target item positions supplied to it by the perceptual field. To actually form peaks, however, both fields need additional node input that elevates the resting level.

The outputs of the reference field and of the target candidates field are subjected to a coordinate transformation (blue diamond in Figure 2.2), which brings the target candidate locations into a coordinate frame that is centered on the location sustained in the reference field. Staying within the neural framework of DFT, the required transformation can be realized as a four-dimensional DF (Schneegans & Schöner, 2012; Lipinski et al., 2012). Transformations were here implemented as convolutions of the two field outputs, however, which is largely equivalent in function but generates less computational overhead when simulating the evolution of activation over time on a computer.

The transformation yields a representation of the positions of potential target items relative to the reference position. This representation is supplied as input to the *relational CoS field* and the *relational CoD field*. The roles of these fields in terms of process organization are analogous to the CoS and CoD color fields, namely detecting a match or non-match of peak positions in the input with the spatial term at hand. The spatial term impacts on the two relational fields through patterned connections from spatial term production nodes (purple circles in the leftmost column in Figure 2.2) to the fields. The connection patterns for each of these projections are displayed at the bottom of the architecture in Figure 2.2 and as larger versions in Figure 2.3.

As these patterns show, an active spatial term production node elevates activation in the relational CoS field in regions that are good matches for

the spatial term for which the node stands, but without pushing activation beyond threshold. A peak emerges only if one of the positions currently represented in the target candidates field lies within the elevated region (the relational CoS field is in a selective regime allowing only one peak). If this occurs, the peak signals that an item matching the spatial term has been found.

Conversely, connectivity of the spatial term production nodes into the relational CoD field instantiates an inverted version of the activation pattern there, elevating regions that do *not* fit the spatial term. Thus, a peak arises in the relational CoD field if a target candidate position lies within the region of non-match. If both a matching and a non-matching target candidate are present, the resulting peak in the relational CoS field homogeneously inhibits the CoD field, preventing the mismatching item from crossing threshold. Thus, a peak in the relational CoD field signals that only items in the desired target color are present that do not match the spatial term at hand.

In case a peak is formed in the relational CoS field, a reverse coordinate transformation (green diamond in Figure 2.2) brings the peak position from the relational CoS field back into the original reference frame of image space. It functions in the same way as the other transformation, but takes input from the reference field and the relational CoS field, each of which contributes a single peak position, and yields a single target location in image coordinates.

The back-transformed location is supplied to the *target response field*, where it results in a peak signifying the location of the visual item which corresponds to the target in the current spatial phrase.

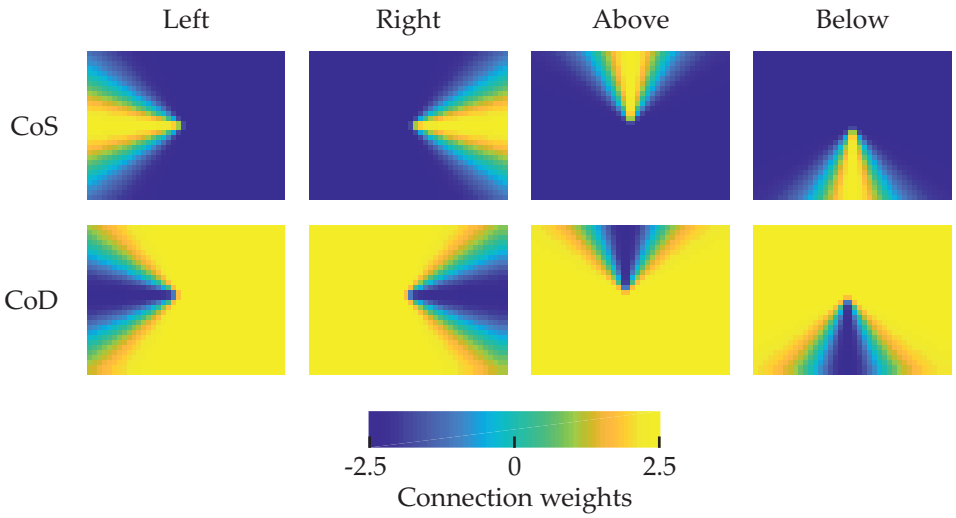


Figure 2.3: Input patterns from spatial term production nodes into the relational CoS field (top row) and the relational CoD field (bottom row).

A connection from the target response field conveys purely spatial input to the perceptual field, resulting in a cylindrical column of sub-threshold activation there, which overlaps with the peak of the target object and thus enhances activation at that location. This is functionally vital for model capabilities not described in detail here, such as responding to "Which item is to the left of the red item?" (see M. Richter, Lins, Schneegans, & Schöner, 2014), that require bringing only the target item into the attentional foreground in order to extract its features.

Finally, a set of substrates ensures that further grounding attempts are performed until successful grounding is achieved, in case previous passes yielded erroneous results. First, whenever an item is stored in the reference field, it passes on the activation to the *reference inhibition of return (IoR) field*, so that a self-sustained peak at the same location is formed there, which in turn feeds back inhibitory input to the reference field. Second, a cascade of control node activations occurs that is hinted at in the bottom right of Figure 2.2. This set of nodes and connections is responsible for resetting activation if no matching target candidate is found for the currently selected reference position in conjunction with the spatial term at hand. In short, a peak in the relational CoD field initiates the cascade by disinhibiting a precondition constraint, thereby allowing the node representing the reset intention (preactivated by input from task nodes, explained in the next section) to become active and briefly push activation in the reference field and the target candidates field below threshold, removing all peaks there, after which the intention is itself switched off by the corresponding node representing the reset CoS. This leads to the start of a new cycle of grounding processes (modulated by peaks in the reference IoR field).

2.2.3 Spatial phrase representation and process organization

As described in Section 1.7, visual spatial relations are characterized by three aspects or roles: a reference item, a target item, and the relation itself. A concrete spatial phrase links a word and thus the associated concept to each of these roles. The current work focuses on the example of color concepts, such as 'red', to define the target item and the reference item, while the spatial term is filled by simple spatial relational concepts, such as 'left of'.

In the architecture, dedicated dynamic nodes signify specific conjunctions of roles and concepts in a spatial phrase. These are shown as blue and purple circles on the left side of Figure 2.2 and are called *memory nodes* (blue) and

production nodes (purple). A pair of these nodes exists for every concept that can fill the target role (e.g., for 'target: red'), as well as for every concept that may be employed as a reference (e.g., 'reference: red'), and for each concept that may fill the role of relation (e.g., 'spatial term: left'). In Figure 2.2, the nodes are organized into rows by concepts and into columns by roles they may fill (note that the control nodes in the top row and described below can similarly be classified as belonging to different roles, as illustrated by their position in the figure).

The memory nodes pose the quasi-linguistic input stage of the model, serving the role of storing a discrete representation of the components of a spatial phrase. A phrase is supplied to the model from outside the system by activating those memory nodes that correspond to the filler-role conjunctions in the phrase (processing real verbal input is beyond the scope of the model). The memory nodes retain this activation due to strong self-excitation.

The production nodes gate the impact of memory node activation on the DF system. A memory node can affect activation in the DFs only through its associated production node, to which it is connected by an excitatory projection (reciprocally, although the backwards connection is relevant only for tasks other than grounding). Through these connections, active memory nodes pre-activate their production nodes, but to become fully active the production nodes must additionally receive input from an intention node (see below).

The production nodes are coupled to different fields by reciprocal, patterned synaptic connections. Color production nodes are connected to different regions of the color intention field (as hinted by small plots of input weights placed above connection lines in Figure 2.2) while spatial term production nodes are connected to different regions in the two relational fields (as already described and according to the connection weight plots in Figure 2.3). Through this connectivity, each production node can evoke a specific pattern of activation in the fields that is an instantiation of the featural or spatial concept the node stands for, within the continuous, metric space the fields represent.

Intention nodes (green nodes labeled 'i' in Figure 2.2) and *CoS nodes* (red nodes labeled 'c') control the sequential order of processing steps that are required for grounding (or other tasks). Most importantly, this pertains to coordinating at which point during processing a concept stored in a memory node is allowed to impact on the fields, as well as at what point which role is filled, and consequently which visual item is assigned to which role. A pair

of an intention node and a CoS node exists for each role. Each of the intention nodes is coupled through an excitatory connection to the corresponding layer of production nodes, so that when the intention node is active, any production node that is pre-activated by a memory node crosses the output threshold and projects an activation pattern into the fields. In addition to that, each of the intention nodes is connected to a specific field that corresponds to one of the roles in a spatial relation (i.e., the reference, target candidates, or relational fields). Through this connection, activation in the respective field is homogeneously elevated, enabling localized inputs from other sources to create peaks there.

The corresponding CoS nodes are pre-activated by input from the intention nodes. As a second input source, the CoS nodes receive input from the role-specific fields. Together, this allows the CoS nodes to signal the formation of activation peaks in the role-specific fields by themselves becoming active (but only when the associated intention is active). This in turn signals the completion of a processing step, such as filling the role of reference by creating a peak in the reference field. In this case, the CoS nodes switch off the associated intention nodes through an inhibitory connection, thus ending the processing step driven by that intention node. The same basic principles govern the function of any triad of intention node, processing substrate, and CoS node.

Processing a spatial phrase that is stored in the pattern of memory node activations is initiated by activating a higher-order *task node* from outside the architecture. This node represents the current demand for a particular cognitive task to be carried out. One such node exists for each type of cognitive task the architecture can perform, with only the one for the task of grounding a spatial phrase in current visual input being relevant here (note that the task nodes are not shown in Figure 2.2). The task node for grounding is connected via excitatory connections to all intention nodes and thus activates all intention nodes that are not inhibited from elsewhere.

Sequentiality of processing steps is generally enforced by *precondition constraints*, which are dynamic nodes (black circles labeled 'p' in Figure 2.2) that inhibit another node until they are themselves inhibited. The most important use of this in the current context is to prevent the target intention node from becoming active before the reference CoS node becomes active, thus enforcing the selection of reference and target object to occur one after the other. This is required since these processes rely on the same perceptual substrates, especially the color and perceptual field, so that simultaneous activation of both

would potentially result in erroneous color-role bindings (expressed through locations stored in the reference and target candidates field).

2.3 Results

This section describes the time courses of the dynamic processes associated with grounding spatial phrases. The results shown here are based on real-time numerical solutions of the differential equations that constitute the model, obtained from its implementation in cedar (Lomp et al., 2016), a software framework for building and simulating DFT architectures.

The spatial phrase that is grounded in both examples below is "The green item to the left of the red item". The employed visual scenes for the first and the second example are shown in panels a and b of Figure 2.1, respectively. Note that although the two examples share the same phrase the grounding processes differ. This is because the reference item is uniquely specified in the first example, as there is only one red item, while it is not immediately clear in the second example which of the two potential reference items the phrase refers to.

Following the two exemplary processing sequences, the evolution of activation in the perceptual field over the grounding process will be considered in more detail.

2.3.1 Grounding a spatial phrase in a visual scene

In the following it is described how the model grounds the phrase "The green item to the left of the red item" within the visual scene shown in Figure 2.1a. The evolution of activation patterns over time for this scenario is shown in Figure 2.4, with the top plot showing the activation of the reference, target, and spatial term intention nodes, and the columns of two-dimensional field plots depicting activation for the time points marked in the node plot. Note that absolute time values are given in Figure 2.4 and 2.5 only for easier reference to individual time points; they represent concrete computation time when simulating the model on a specific computer and thus do not correspond to the amount of time the same processes might take in a real neural system. The two-dimensional activation landscapes provided for the perceptual field in Figure 2.4, 2.5, and 2.6 show the maximum of field activation along the color dimension.

Before processing of the grounding task is initiated, the spatial phrase

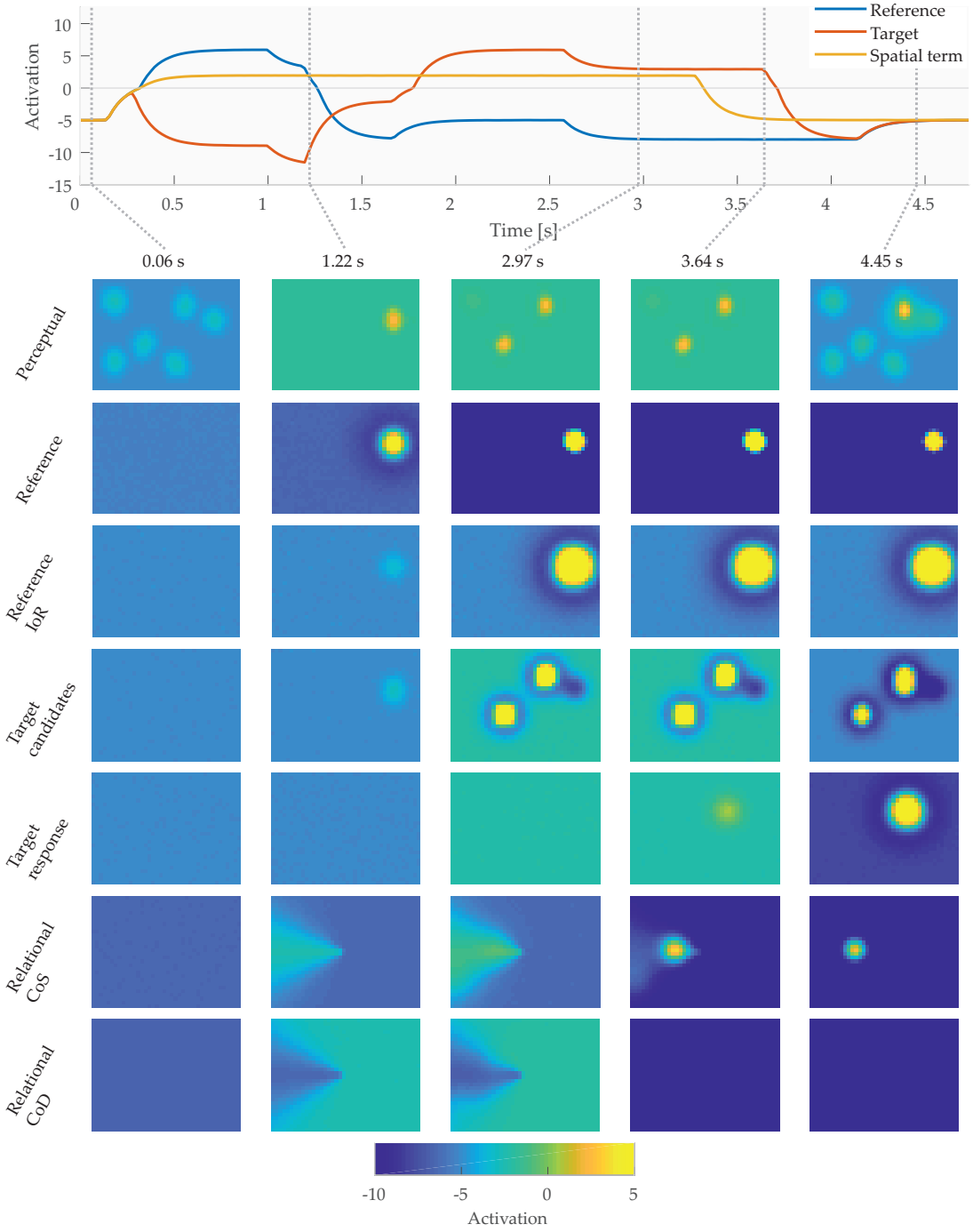


Figure 2.4: Evolution of activation patterns in the model while it grounds the spatial phrase "The green item to the left of the red item" in the scene in Figure 2.1a (one reference candidate). Activation time courses are shown for the intention nodes (top), and activation patterns of relevant fields are shown for five selected points in time (bottom). Field activation is color coded according to the color bar at the bottom. The two-dimensional activation plots of the perceptual field show the maximum of its activation along the color dimension. See text for details.

is encoded as an activation pattern in the memory nodes by activating the memory nodes 'reference: red', 'target: green', and 'spatial term: left of'. The field plots at 0.06 seconds in Figure 2.4 show the state of the model before the grounding task has started. All fields and intention nodes are at the resting level, except for the perceptual field, which displays a sub-threshold pattern of activation caused by the visual items in the input scene. Task processing is initiated at about 0.1 seconds by activating the grounding task node. The remaining task is performed by the architecture autonomously, without external control.

As the grounding task node is activated it sends activation to all intention and precondition nodes. As the node plot in Figure 2.4 shows, this results in the reference intention node and the spatial term intention node becoming active, whereas the target intention node is inhibited by the precondition node and is thus depressed below its resting level for the moment.

The now active spatial term intention node supplies activation to all spatial term production nodes. The production node for the spatial term 'left of' also receives input from the corresponding memory node and thus becomes fully active. It projects its spatial template (see Figure 2.3) into the relational CoS field and the reversed pattern into the relational CoD field, both resulting in a sub-threshold activation landscape (Figure 2.4, snapshot at 1.22 seconds).

At the same time, the now active reference intention node sends activation to all reference production nodes. Since the one for 'reference: red' is pre-activated by its memory node, it crosses the output threshold and projects localized Gaussian input into the color intention field. This input brings activation in the color intention field above threshold at the field site corresponding to the color red.

The peak in the color intention field in turn produces output that impacts on the perceptual field, giving rise to a slice of elevated activity homogeneous over space but local in color space, centered on the color red. The slice enhances the activation bump in the perceptual field that signifies the single red item in the input, causing it to develop into a peak of activation (snapshot at 1.22 seconds). This activation is relayed to the reference field. Because activation in that field is slightly elevated by the active reference intention node, the location of the red reference item is stored in the reference field (snapshot at 1.22 seconds). Note that the target candidates field receives the output from the perceptual field as well but cannot form a peak since it is not pre-activated by the target intention node.

As a consequence of the peak forming in the reference field, the reference

CoS node is activated, signalling that a reference item has been chosen. The reference CoS node inhibits the reference intention node, causing it to become inactive at about 1.3 seconds. It also inhibits the precondition constraint node that previously inhibited the target intention node.

In consequence, the target intention node becomes active (due to being excited by the grounding task node), crossing the output threshold at about 1.8 seconds. It sends activation to all target production nodes. Since only the production node 'target: green' is pre-activated by its memory node, only this node is activated and projects an activation pattern corresponding to the color green into the color intention field. As before, the ensuing peak in the color intention field creates a slice of activation in the perceptual field, this time enhancing activation for all bumps created by green items. Thus peaks arise in the perceptual field at the two positions where green items are located (snapshot at 2.97 seconds).

The output of the perceptual field is relayed to the target candidates field, which can now form peaks due to being homogeneously pre-activated by the target intention node. Thus, the two potential target locations are stored as peaks in the target candidates field (snapshot at 2.97 seconds).

Note that, in the meantime, reference field output has lead to a peak in the reference IoR field as well (snapshot at 2.97 seconds), which in turn inhibits the reference field at the same location (but not strong enough to remove the existing peak there).

Now that peaks exist in both the reference field and the target candidates field, peak locations in the target candidates field are transformed to a coordinate frame centered on the reference item, and the output of the transformation is supplied to both relational fields. This is very slightly visible in the snapshot at 2.97 seconds, in the form of a developing bump of activation to the left of the center of the relational CoS field (the other target candidate is discernible only through slight elevation of activation in the lower left corner).

In the snapshot at 3.64 seconds, a full-fledged peak has built in the relational CoS field, since the input from the target candidates field that corresponds to the upper green item overlaps with the region of elevated activation caused by the projection of the spatial term production node. This supra-threshold activation on the one hand inhibits the relational CoD field, preventing the non-matching item from creating a peak there, and on the other hand activates the spatial term CoS node. This ultimately inhibits the spatial term intention node, removing the spatial template input to the relational fields.

Furthermore, now that a peak exists in both the relational CoS field and the reference field, the backwards transformation becomes active, bringing the peak in the relational CoS field back into image coordinates and projecting it into the target response field. The latter builds a peak at the location of the final target item, which is already discernible in the snapshot at 3.64 seconds and has fully developed in the snapshot at 4.45 seconds.

The output generated by the peak in the target response field is sent to the target CoS node, activating it, and in consequence switching off the target intention node. This removes the slice of activation in the perceptual field so that all peaks there vanish. Also, the peak in the target response field projects a column of activation into the perceptual field, thereby re-establishing the peak that corresponds to the target item (snapshot at 4.45 seconds). Thus, the correct target object has been located and attentionally selected.

2.3.2 Grounding with multiple reference candidates

In the above example the reference item was unique: Only one red item existed in the scene. Thus, lacking an alternative, the correct reference item was attentionally selected and stored in the reference field immediately. How the model solves the case where multiple items share the reference-defining feature is described here. The scene in Figure 2.1b was used as input to the model in conjunction with the same spatial phrase as before ("The green item to the left of the red item"). The evolution of activation for this scenario is shown in Figure 2.5.

As apparent when comparing Figure 2.5 to Figure 2.4, the first half of the processing sequence in the current example is very similar to the previous one. First (up to the snapshot at 1.21 seconds), reference and spatial term intention nodes become active, so that all potential reference items are brought into the attentional foreground in the perceptual field. The difference to the first example here is that it is not clear which of the two red items should be stored in the reference field. The one on the lower left is selected, chosen over the other one based on random noise or slightly differing saliency of the two items, and its position is stored in the reference field.

In the snapshot at 3.39 seconds, the reference intention node has been inhibited by the reference CoS node and the target intention node has become active instead. The resulting slice of activation in the perceptual field at the value of the target color (green) has brought above threshold the two target candidates, and their positions have been stored in the target candidates field.

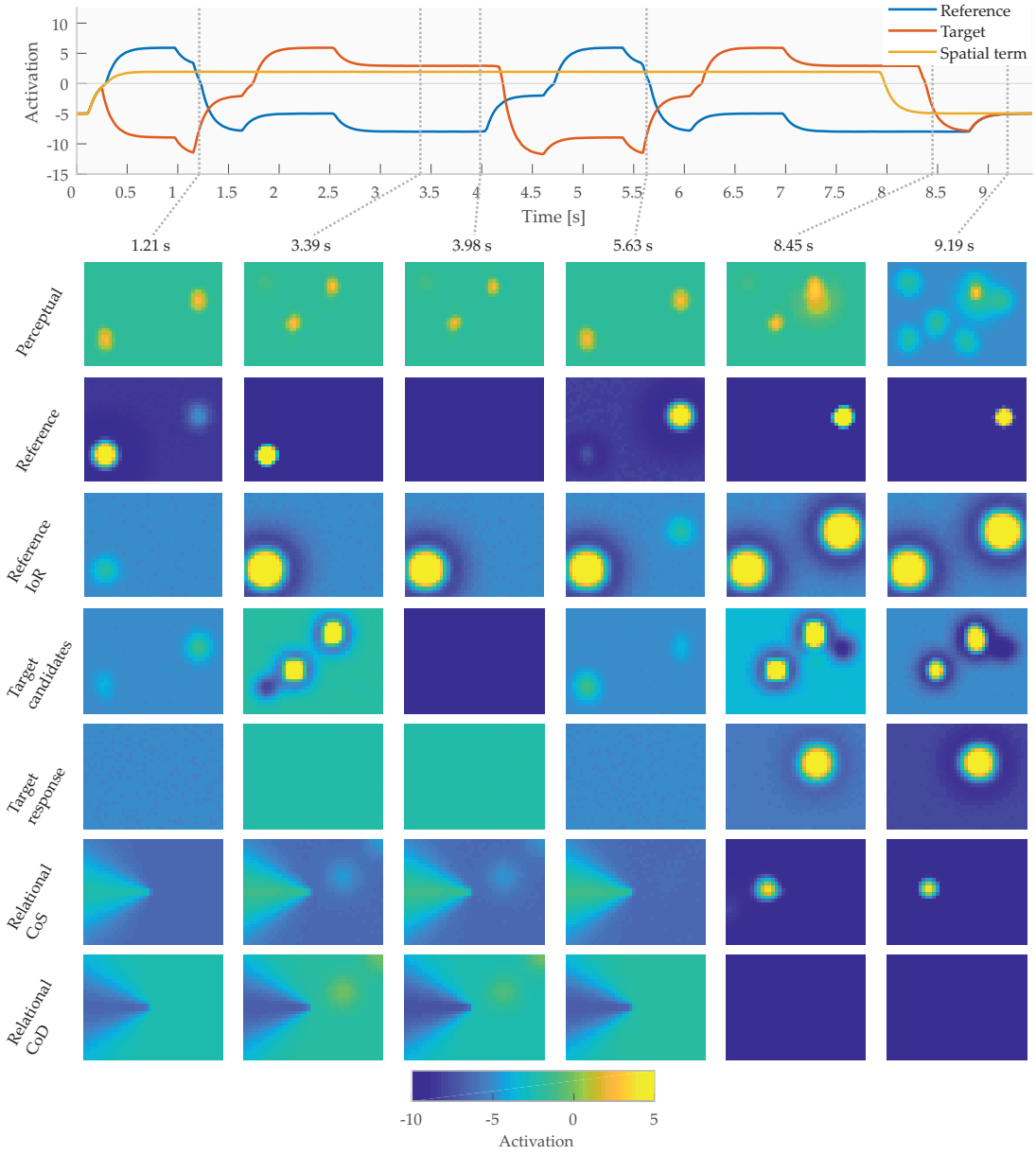


Figure 2.5: Evolution of activation patterns in the model while it grounds the spatial phrase "The green item to the left of the red item" in the scene in Figure 2.1b (two reference candidates). Activation time courses are shown for the intention nodes (top), and activation patterns of relevant fields are shown for six selected points in time (bottom). Field activation is color coded according to the color bar at the bottom. The two-dimensional activation plots of the perceptual field show the maximum of its activation along the color dimension. See text for details.

Also, a peak at the location of the currently selected reference position has developed in the reference IoR field.

With peaks in both the target candidate field and the reference field, the coordinate transformation brings the target candidates into a frame centered on the currently selected reference position and projects it into the relational CoS and CoD field. In contrast to the example in Figure 2.4, however, none of the target candidates matches the spatial term in relation to the currently selected reference position. Therefore, none of the target candidate positions overlaps with the region of elevated activation in the relational CoS field, so that no peak is established there. As the relational CoS field does not generate output, the relational CoD field is not inhibited. Because the target candidate positions overlap with the reverse activation pattern there, peaks arise in the relational CoD field, slightly visible in the snapshots at 3.39 and 3.98 seconds.

This signals that none of the target candidates fit the spatial term at hand when it is applied to the currently selected reference position. The output generated by the peaks in the relational CoD field is projected through an inhibitory connection onto the precondition constraint node shown at the bottom right in Figure 2.2, which initiates a cascade of activations in the nodes shown in its vicinity. The precondition constraint node is depressed below threshold, so that it no longer inhibits the reset intention node. Being activated by the grounding task node, the reset intention node becomes active and depresses activation in the reference field and the target candidates field through inhibitory projections, deleting all peaks there (snapshot at 3.98 seconds). The reset intention node is deactivated shortly after by the respective CoS node. Note that the peak in the reference IoR field is retained (snapshot at 3.98 seconds).

With no peak remaining in the reference field, the reference CoS node becomes inactive which, first, allows the reference intention to become active again and, second, reinstates activation of the precondition constraint which in turn deactivates the target intention node (both happens at around 4.5 seconds).

The resulting activation state of the architecture is similar to the state before the start of processing and allows a new attempt at grounding to be carried out. However, the peak in the reference IoR field is carried over from the first pass and affects reference selection in the second pass. This occurs between the snapshots at 3.98 and 5.63 seconds: Through locally inhibiting the reference field, the reference IoR field prevents the lower left red item to be selected a second time. As shown in the snapshot at 5.63 seconds, both red

reference candidates are above threshold in the perceptual field, but this time the upper right item's position is stored in the reference field.

The remaining processing sequence follows the same course as the example in the previous section, so that ultimately the correct target item is selected in the perceptual field (snapshot at 9.19 seconds).

Note that the case described so far, in which the incorrect reference item is selected first, is only one of two possible sequences that may occur in response to the scenario. If the correct reference item is instead selected in the first pass, the processing sequence is shorter and equivalent to the second half of the sequence just described.

2.3.3 Evolution of activation in the perceptual field

The perceptual field is the locus in the model that reflects most directly the momentary relevance of different visual items for the task at hand. This makes it a possible source of guidance for motor responses directed at task relevant items, such as those in the experimental task described later (see Chapter 3). It is therefore considered in more detail here how activation evolves in the perceptual field and which item locations are active above threshold in the course of grounding a spatial phrase.

Figure 2.6 compares the evolution of activation in the perceptual and reference field between the different scenarios. Letters to the right of each column of field plots indicate which items are above threshold in the perceptual field in each snapshot. R and R_A refer to the 'correct' reference item, R_B refers to the item sharing the reference color, T refers to the target of the spatial phrase, and D refers to the distractor item, that is, the item sharing the target color but providing a worse match to the spatial term than the target itself (as per the weight patterns in Figure 2.3).

The scenario in Figure 2.6a is identical with the first example described above (see *Grounding a spatial phrase in a visual scene*, p.49). The reference item is selected first (due to the precondition constraint inhibiting the target intention node), which in this case means that the first peak in the perceptual field is located at the position of the single reference item. After storing the reference location, a short period without supra-threshold activation in the perceptual field follows, during which the target intention node becomes active. Next, all items in target color form peaks in the perceptual field, including the distractor and the target item. They remain above threshold longer than the two reference candidates before, due to the multiple processes taking place to

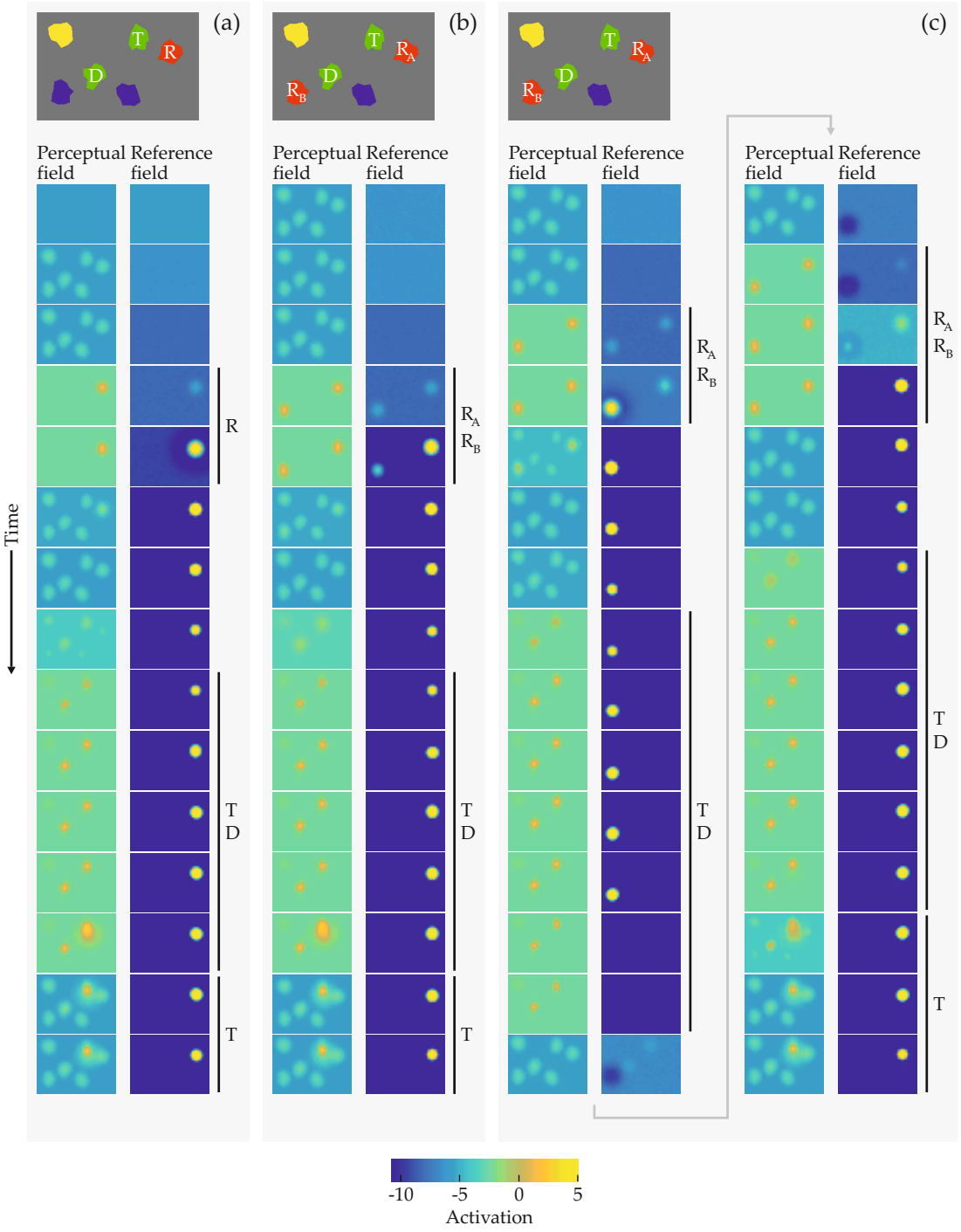


Figure 2.6: Evolution of activation in the perceptual field and in the reference field for grounding the spatial phrase "The green item to the left of the red item" in the visual scenes at the top of panels a, b, and c. Visual items are labeled according to their role in the scenario, that is, as target (T), distractor (D), reference (R or R_A), or sharing reference color (R_B). Black bars in the bottom part labeled in the same way indicate which item locations are above threshold in the perceptual field in the corresponding snapshots. Consecutive snapshots are separated by 320 millisecond steps.

evaluate their match with the spatial term (two transformations and matching item position to the spatial term in the relational fields). Finally, the target item is identified as the best match for the spatial term and remains the only item above threshold in the perceptual field.

Figure 2.6b shows the case not covered in detail so far, in which multiple items share the reference color but the correct one is selected in the first pass. In terms of computational steps, this scenario progresses in the same way as the previous one. However, an important difference is that during the step of reference selection both items in reference color are brought above threshold, not just one. This is mandatory, even though R_B is never explicitly used for evaluating the relation, in order make explicit the set of possible reference items and to be able to choose one item from that restricted set. Then both the target and the distractor become active again, after which the final target is selected in the perceptual field.

Figure 2.6c shows the same scenario as (b), but with the incorrect reference item being selected first (as in the second example described above; see *Grounding with multiple reference candidates*, p.53). The first part of processing is the same as in panel b, that is, first both reference candidates are active in the perceptual field to select a (tentative) reference location from among them, and then both target candidates (target and distractor item) are brought above threshold to assess spatial term match. Since none of the target candidates matches the term, however, the processing sequence is restarted in this example. The ensuing processes again involve bringing above threshold the reference candidates and the target candidates, after which only the correct target item is selected in the perceptual field.

2.4 Discussion

This chapter demonstrated how a pervasively neural system can ground language in perception. The grounding task was achieved through controlled interaction of modal representations of sensory experience with amodal, discrete representations of linguistic information. Concretely, dynamic nodes received the components of a spatial phrase as input, while visual input was supplied to dynamic fields. The two types of representation were linked through neural connections such that node activation guided the selection of spatial phrase referents in the dynamic fields. The progression through the processing sequence was autonomously controlled by a layer of neural structures likewise implemented as nodes and fields. As a result, the correct

referents were ultimately found in the visual scene.

Concretely, it was demonstrated how the model grounds the spatial phrase "The green item to the left of the red item" in two different visual scenes. There were two green items in each of these scenes, and in each case the model ultimately selected the one best matching the spatial term as final target, bringing that item above threshold in the perceptual field. To determine the final target, the overlap of all target candidates with a spatial template was assessed by projecting both to a spatial representation centered on the reference item. Successful completion of the steps of reference selection, target selection, and of finding a pair matching the spatial phrase was signaled by neural conditions of satisfaction.

In one scene, there was only one item of the color that filled the role of reference in the spatial phrase. Determining the reference item referred to in the phrase was straightforward in this case, requiring only one attempt of grounding. In the other scene, two items shared the reference color, which potentially required to test different hypotheses about which of these reference candidates the phrase referred to. This was achieved through a neural condition of dissatisfaction being triggered by target candidates not matching the spatial template when the incorrect reference candidate was selected first.

In all scenarios, the selection of reference and target items occurred in two separate steps, as a consequence of shared substrates being recruited to link color to location. In each step, all items sharing the respective color were brought above threshold in the perceptual field, in order to select one of them through a selection decision. The sequence of items being represented in the perceptual field during the grounding process was shown to differ between the scenarios and depended on the selection order of reference candidates.

Note that the order of selection in the model — first reference, then target item — is arbitrary insofar as trivial modifications to coordinate transformations and spatial templates would allow the reverse order to be used (in which case multiple items sharing the target feature would potentially lead to multiple passes being necessary; M. Richter et al., 2017). Experimental evidence in this respect is similarly inconclusive (Yuan et al., 2016; Franconeri et al., 2012; Burigo & Knoeferle, 2015; see *Evaluating spatial relations*, p.22).

In summary, the model realizes the essential steps postulated by Logan & Sadler (1996) in an autonomous manner, guided by a spatial phrase. Aspects not addressed include the selection of reference frames (e.g., Carlson-Radvansky & Irwin, 1993; Schultheis, 2007), and the impact of functional object properties on how spatial templates are applied (e.g., Carlson-Radvansky

et al., 1999). Both are outside the scope of the model, as only a viewer-centered reference frame with varying origin was used here and object representation was impoverished.

However, the main purpose of the model was not to capture all aspects of how humans process spatial relations, but to devise a prototype mechanism for grounding language within a framework of neural plausibility and embodiment, with spatial relations as a test case. The outcome of this approach is used as a qualitative heuristic to derive and interpret possible effects of the task's embodiment on motor behavior, rather than making meticulous quantitative predictions or fitting narrowly circumscribed experimental evidence.

To this end, multiple constraints of different nature were combined in the model. Some of these derived from the framework within which it was developed (i.e., DFT and the associated stance of embodiment). Only biologically plausible neural concepts from the framework of DFT were employed. Other constraints derived from functional demands of the task itself. Autonomous process organization must give rise to ordered steps, which is a specific challenge for fully neural systems and therefore required substrates in addition to those directly linked to sensorimotor surfaces. Finally, basic aspects of how the task was solved were derived from experimental evidence about the sequential attention shifts required to process visual relations (Franconeri et al., 2012; Yuan et al., 2016; Burigo & Knoeferle, 2015; Logan, 1994), which are in line with the fundamental attentional limitations in visual perception (Schneegans, 2016; Schneegans, Spencer, & Schöner, 2015; Treisman & Gelade, 1980; Hyun et al., 2009; Wolfe, 1994).

The combination of these sources of constraints is unusual for cognitive models in the domain of higher cognitive function, particularly the commitment to refrain from non-neural placeholders to connect neural components and to account for all aspects of process organization in a neural manner. The model thus complements rather than refutes the merits of approaches to spatial language grounding that do not adhere as strictly to these constraints. This, however, also tends to make direct comparisons between these different approaches difficult.

In a model by Vavrečka & Farkaš (2014), for instance, unsupervised learning methods (e.g., self-organizing maps) allow the system to acquire links between linguistic sentences and spatial relations of two items in visual input, by binding both sources of input in a conjunctive representation. In one variant, the model employs a position-invariant "what" system representing color and shape, and a feature-invariant "where" system representing item

location, which converge on a multimodal representation that also receives input from a phonological representation of linguistic phrases (with fixed grammar). The two visual items are foveated by default, however, and it is not considered how the selection of items from the visual input is organized under attentional constraints. Generally, this approach explored how different types of representations and unsupervised methods may be used in grounding relations rather than considering the organization of information flow through the system.

Schultheis & Barkowsky (2011) presented a computational model of spatial reasoning that integrates discrete long-term memory structures, space-analog representations, visual inspection processes, and distributed control of the interaction between the system's components. While committed to linking to human cognition, however, the architecture mainly draws on concepts of non-neural information processing.

In a different spirit, Regier & Carlson (2001) proposed a widely recognized mechanism for assessing the match of a target object to a spatial term relative to a reference object. The mechanism is based on describing the relation between a reference object and a target object in the form of a vector sum weighted by attention directed at the objects. The mechanism is neurally based in the sense of being inspired by neural population vectors that have captured arm movement direction based on motor cortical population activation (Georgopoulos et al., 1986). The steps required to perform the computation, however, are not embedded in a neural system that would implement and control the different aspects of the task.

The model described in this chapter is embedded within a larger framework of DFT modelling efforts that have addressed related questions. These models are related not only conceptually but also share similar neural substrates with overlapping functionality. The perceptual field in particular is analogous to a set of fields in previous DF models of visual scene representation (Schneegans, Spencer, & Schöner, 2015; Schneegans, 2016; Zibner et al., 2011; Zibner, 2017). Similar to the perceptual field here, these fields interface with visual input and relay or gate the impact of visual items via attentional selection, thereby providing the basis for forming and updating visual working memory, as well as detecting changes in item features, positions, or conjunctions thereof, and for providing input to object recognition. Thus, DFs that occupy similar functional roles as the perceptual field here may be coupled to different cognitive systems, in line with the view that neural substrates are typically implicated in various functional domains (see *Embodiment*

in DFT, p.10).

In light of the later described experimental assessment of computer mouse trajectories (see Chapter 3) within an experimental paradigm analogous to the task solved by the model, a possible next step is to extend the model to generate such movement trajectories. In principle, this is possible, as demonstrated by DFT-based work in which robotic arm movements were guided by field activation (Tekülve et al., 2016; Zibner et al., 2015). However, as exemplified by these studies, a number of non-trivial issues is associated with the generation of biologically plausible end-effector movements in an embodied system, such as transformations between retinal and end-effector coordinates, coupling to a muscle model, and the impact of visual feedback on arm motion. These issues make such an endeavor a separate research program rather than a trivial model extension (but see Lepora & Pezzulo, 2015, for a model that aims to link perceptual decision making to motor action in a less process-oriented manner).

Therefore, for now, the link between model and experiments must be drawn by the expectation that the processes operating on the grounded substrates will have an impact on the motor level. That this may be the case is suggested by the various types of evidence reviewed initially (see *Embodiment in DFT*, p.10, and *Probing embodiment with mouse tracking*, p.16), in particular the fact that dorsal stream and motor cortical activation patterns vary with task characteristics (Cisek, 2007; Cisek & Kalaska, 2005; Bastian et al., 1998). This hinges on the assumption, of course, that selection processes in the model are comparable to the factors that biased activation in these studies.

One aspect of the grounding process in the model may be of particular relevance with respect to potential motor signatures. Namely, the model suggests that in the two steps where target and reference items are selected, all items that share the color of the respective item are brought above threshold in the perceptual field at the same time. This occurs even in cases where the correct relational pair is found and grounded in the first attempt. This makes it more likely that signatures of all items that are potential role-fillers for the spatial phrase (by virtue of their features) may be found in behavior.

As described initially, previous evidence similarly suggests that items must be spatially indexed (and that this occurs sequentially). In the model, this selection step is made concrete as a dynamic mechanism for feature-based search that affects all locations in a spatial representation at the same time, as seen in visual search experiments (e.g., Treisman & Gelade, 1980; such a mechanism may also apply if a target is defined by multiple features; Wolfe,

1994). This mechanism leads to the simultaneous activation of all items sharing the feature of the item currently being selected (target or reference).

The simultaneous activation is also required by the mechanism of the model that matches target item positions to spatial templates. Activating the positions of all target candidates at the same time allows to transform them into the reference-centered frame in parallel and match them to the spatial template simultaneously.

Transforming all potential target items at the same time is not directly supported by experimental evidence, mainly because research about relation processing has focused on distractor items that are not identical to the target (e.g., Logan & Compton, 1996). However, the mechanism poses the most efficient mechanism that is possible under the constraints that shaped the model (particularly limited-capacity binding via space and the neural restrictions posed by DFT). Selecting only one reference and one target candidate in each attempted grounding pass would result in a combinatorial number of item pairs to be tested until the pair is found that matches the spatial relation. As a further complication, to select the best-matching target item instead of a less than ideal candidate, it would be required to store match quality for each tested pair and compare different pairs on this dimension. Also, in order to not test the same pair repeatedly, a representation of inhibition of return for item *pairs* would be required (which, as a field, would span four dimensions).

Other manners of processing are possible under relaxed constraints. For instance, high-dimensional neural representations in which language, relations, and features are integrated (e.g., Vavrečka & Farkaš, 2014) may solve the task more efficiently, but run counter to the constraints of the framework of DFT committed to here.

Chapter 3

Behavioral Signatures of Embodied Spatial Language Grounding

The experiments described in this chapter aimed to show that cognitive processes of spatial language grounding operate on modal substrates that are firmly embedded in the sensory-motor loop. To this end, motor responses directed at the visual targets of spatial phrases were examined for biases associated with the grounding process.

The general task was the same in all experiments: Participants saw a spatial phrase, such as "The green item to the left of the red one", followed by a visual display with multiple colored items, and moved the mouse cursor onto the target item denoted by the phrase. Mouse trajectories were recorded during the task.

Behavioral signatures of grounding processes were expected to take the form of attraction toward visual items that, based on their color, were potential fillers for one of the roles in the spatial phrase (reference or target). This expectation stemmed from the fact that these items must be attentionally selected during relation grounding, as realized in the model described in the previous chapter.

A motor impact of this selection was expected due to the postulate that processes during grounding operate on substrates closely linked and similar in structure to sensorimotor surfaces, as established in the section *Embodiment in DFT (p.10)*. That task characteristics may impact motor responses is also suggested by the evidence described in the section *Probing embodiment with mouse tracking (p.16)*. This evidence established that motor responses are affected by the certainty over solutions in abstract cognitive tasks as well as by uncertainty about movement targets induced through delayed cuing of the target. The processes of language grounding were expected to bias motor

decision making in a similar way as task variables in these earlier studies.

Seven experiments were conducted:

- Experiment one looked for attraction toward items implicated in the grounding process, namely a distractor item that shared the color of the target, and the reference item. In addition, it examined an effect of the spatial term that was more similar to classical embodiment effects in which language biases motor action in accord with implied directionality .
- Experiment two generalized the findings of experiment one to different response metrics, namely horizontal instead of vertical mouse movement. This also served to disambiguate the nature of two effects observed in experiment one.
- Experiment three examined whether the same effects could be observed with a higher gain in mouse cursor movement, more similar to that used in previous mouse tracking studies.
- Experiment four explored the impact of word order in the spatial phrase on the effects observed in the previous experiments and otherwise posed a replication of experiment one with a higher number of participants.
- Experiment five was equivalent to experiment four apart from using a horizontal instead of a vertical response direction. It thus likewise explored the effect of word order and posed a replication of experiment two.
- Experiment six probed whether attraction caused by a competing relational pair transcended the sum of biases evoked by individual items that were not part of such a pair. This was done to gain further support for the hypothesis that attraction effects observed in the experiments here were signatures of flexibly combined grounding processes rather than resulting from stereotypical contributions of individual items.
- Experiment seven sought to provide further support for the interpretation of experiment six in terms of additional attraction being caused by the combination of items into a relational pair rather than generic interaction between multiple items.

For each of the seven experiments, the specific goals, rationales, procedures, materials, and results will be separately described and briefly discussed; a general discussion of the experimental results then follows. Since

the overall paradigm is the same for all experiments, however, most key methodical aspects will be included in the section of experiment one and omitted thereafter. Subsequent experiment descriptions will build on preceding ones and cover only extensions or differences to what has been described before.

3.1 Experiment one: Effects of distractor, reference, and spatial term[1]

The main focus in the first experiment was whether the attentional selection of items that are not behavioral targets but potential role fillers for the spatial phrase would be visible in mouse trajectories. This included the reference item and a distracter item which shared the color of the target item but provided a worse match for the spatial term.

In experiment one, the visual displays in which the spatial phrases had to be grounded contained one item in the reference color (i.e., the *reference* item) and two items in the color of the target specified in the phrase. One of the latter two items provided a better match for the spatial term than the other one. The better-matching item was expected to be selected by the participants and will be referred to as the *target* item, while the other one will be referred to as the *distractor* item. The only difference between the target and the distractor was how well they matched the described relation. Participants were not explicitly told that there would be a target and a distractor item, but were only instructed to select the item best-described by the spatial phrase preceding the display. Therefore, the target and the distractor were viewed as potential movement goals that must be disambiguated through grounding the spatial phrase.

As stated above, the reference item, which was involved in solving this task, was unique in color in each stimulus display. Because only six clearly distinguishable colors were used in each display, and since visual search for sufficiently different colors is efficient (Wolfe & Horowitz, 2004; Wolfe et al., 1990), it was assumed that participants would not deem the reference item a potential movement target from early on.

[1]Parts of the material in this section, including results, have been published in Lins & Schöner (2017), representing joint work of the current work's author (JL) and Gregor Schöner (GS). JL designed, implemented, and conducted the experiment and the associated data analyses. GS and JL engaged in conceptual discussion of the experimental paradigm in the development process. Note that minor differences to the results reported in the original publication may arise due to adjustments in analysis methods.

Apart from target, distractor, and reference item, each display was populated with nine differently colored *filler* items. This was done to prevent solution strategies based on the overall gestalt of the array of reference, target, and distractor, other than grounding the relational phrase. Filler items were expected to not systematically impact the grounding process itself, again due to the ease with which relevant items can be singled out through visual search based on color. This expectation is also supported by a study in which the impeding effect of distracting items irrelevant to a sought relation disappeared when target and reference were colored differently than the other items (Logan & Compton, 1996).

The distractor was hypothesized to metrically attract the trajectories, in analogy to the effect of alternate but incorrect choice alternatives in classical mouse tracking research. However, an important difference was that, here, the spatial location of response alternatives was not known in advance, and it was presumed that the task was solved not in an abstract form but directly within sensorimotor representations of visual space.

The reference item was as well hypothesized to attract mouse trajectories. This effect was expected to result not from target uncertainty, given that the reference was no eligible movement target, but from the allocation of attention to the reference item in the course of selecting it for spatial language grounding.

Finally, a bias into the direction described by the spatial term was hypothesized to occur. This bias was expected to be not directly connected to the grounding process, but to result from a motor priming of the movement direction consistent with the spatial term. A similar effect has been shown in previous research where mouse trajectories were biased by subliminally presented directional prime words (Tower-Richardi et al., 2012). Due to its presumed independence from the visual arrays, the spatial term effect was expected to occur earlier than item-based biases, and to start with movement onset.

3.1.1 Methods

Participants

Twelve participants (five female, seven male) with a mean age of 27.4 years (SD = 3.8 years) were recruited by notices around the local campus, signed informed consent (Appendix B), and received €10 for participation. All but one participant were right-handed, as determined by the Edinburgh Handedness

Inventory (Oldfield, 1971). The participants were naïve to the experimental hypotheses, native German speakers, had self-reported normal or corrected-to-normal vision, and no color vision deficiencies (assessed with the questionnaire in Appendix A).

Procedure

Each trial began with a black start marker (diameter 6.8 mm/0.56° of visual angle, v.a.) in the bottom center of an otherwise gray screen (Figure 3.1a). To proceed, the participant moved the mouse cursor (a white dot with diameter 5.6 mm/0.45° v.a.) onto the start marker, upon which detection of the cursor was indicated by a black ring around the start marker (Figure 3.1b).

After resting on the start marker for 300 ms, a German spatial phrase appeared (Figure 3.1c), centered on a position somewhat random around the center of the stimulus region (up to ±48 mm/20 mm horizontically/vertically; the text was in Arial with a height of 8.8 mm). The spatial phrase could read, for instance, "Das Rote rechts vom Grünen." (translating to "The red one to the right of the green one."; see *Spatial phrases*, p.72). It thus denoted the target item by a combination of a color ("red") and a position given relative to the reference item ("right of"). The reference item was specified only by its color ("green"). The display duration of the phrase varied randomly from one to two seconds, in order to counteract anticipatory responses. If mouse movement occurred while the phrase was still visible, the trial would be aborted and appended at the end of the trial list in order to be presented again later. In this case, a two second feedback was shown prompting the participant to start moving only after the phrase was removed.

Else, the phrase disappeared and simultaneously a beep signaled the participant to start moving the cursor upward. Movement had to be started within one second after phrase offset (Figure 3.1d) or the trial would be aborted and appended at the end of the trial list in order to be presented again later. In this case, a two second feedback was shown indicating that the movement had been started too late. The intent behind this time limit was to standardize the time participants had for forming a mental representation of the described relation. Movement onset was registered when the mouse pointer exceeded a velocity of 20 mm/s. Presenting task stimuli only after movement onset produces more consistent deviation than showing stimuli first (Scherbaum & Kieslich, 2017).

At movement onset, twelve colored items appeared above the start marker (Figure 3.1e). This means that mouse movement was already in progress

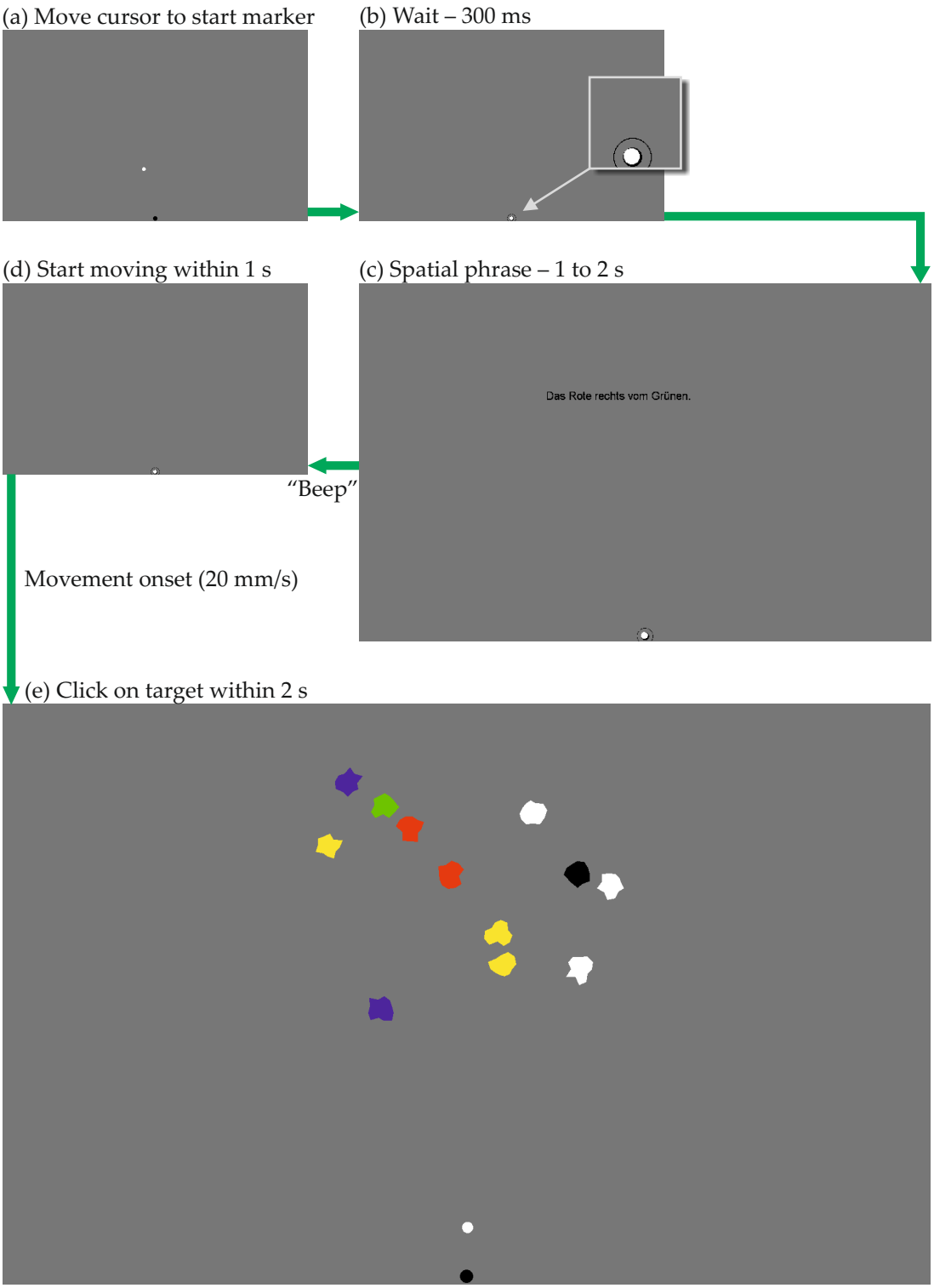


Figure 3.1: Course of events in a single trial.

when the visual items appeared. One of the twelve items was the uniquely colored reference item mentioned in the spatial phrase, one was the target item, and one was the distractor item, defined by sharing a color with the target but providing a worse match for the spatial term in the phrase (according to a spatial template described in *Assessing spatial term fit*, p.73).

The participant's task was to select the item which in his or her opinion best matched the preceding phrase (participants could select any item). Starting with movement onset, participants had two seconds to select an item by clicking within its outer radius (defined in *Visual displays*, p.72). If no item was clicked within two seconds, the trial was aborted and appended at the bottom of the trial list in order to be repeated later. In this case, a two second feedback was shown indicating that the maximum time had been exceeded. The time limit served to prevent participants from stopping mouse movement while grounding the relation, so as to time-lock movement onset and the start of relational processing. The allowed duration was based on pilot work and adjusted to impose a sense of time pressure without requiring hasty responses. Trials exceeding the time limit mainly occurred during the first few trials, before participants adapted to the paradigm, and as mean movement times showed, the two second limit posed a relatively liberal threshold. As soon as an item was selected, the items disappeared and the next trial began, proceeding in the same manner.

Participants were instructed to select the item that they thought provided the best match to the spatial phrase shown before the visual display. They were told that there were no correct or incorrect responses, but to not base their decision on which item was more convenient to reach. Furthermore, participants were instructed that the items did not pose obstacles for mouse movement, that response time was limited such that they had to respond promptly but not hastily, that they could rest in between trials if fatigued, and that they should not lift the mouse off the desk while responding, but that they could do so if necessary for repositioning the mouse in between trials (i.e., during the phase shown in Figure 3.1a).

Before the experimental trials were presented, the experimenter demonstrated the procedure by completing two trials (once choosing the distractor and once choosing the target) and each participant completed 13 practice trials without any time limits. After that, each participant completed 446 trials in random order (one completed eight more, to use the entire set of 5360 trials described in *Generating visual displays*, p.75).

Material

The experiment was implemented and run using MATLAB R2017a and the Psychophysics Toolbox 3 (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). Spatial phrases and visual displays were presented on a 22" LCD screen (Samsung, 226BW at 1920 × 1080 resolution; size of visible image 475 mm × 297 mm) at a viewing distance of approximately 70 cm (thus subtending approximately 40.4° × 22.99° v.a.). Trajectories were collected using a standard computer mouse (Logitech, M-UAE96, mean sampling rate was approximately 92 Hz; note that starting from experiment two a Roccat Kone Pure mouse was used, with an effective mean sampling rate of approximately 400 Hz). Mouse speed was set such that movement on the tabletop translated to cursor movement over the same distance on the screen, in order to make motions more similar to natural arm movements and simplify cognitive transformation from hand coordinates to screen space (see, e.g., Krakauer et al., 2000).

Spatial phrases Spatial phrases were in German and of the form exemplified by "Das Grüne rechts vom Roten.", translating to "The green [one] to the right of [the] red [one]". That is, each phrase started with the article "Das", followed by a nominalized color word specifying the target item of the trial ("Grüne"), a spatial term specifying a spatial relation ("rechts vom"; where "vom" is a contracted form of preposition and article), and another nominalized color word specifying the reference item of the trial ("Roten.").

Table 3.1 lists the candidate words, labeled source sets, from which the spatial phrases were constructed such that they corresponded to the item arrangements in the visual display of the trial at hand. In all trials, the spatial phrase posed a valid description of an item in the stimulus display.

Visual displays The general configuration of visual displays on the screen is shown in Figure 3.2. The item arrays were made up of irregular polygons consisting of 12 vertices placed around the item center in equal angular intervals and at random distances between 8.2 mm (0.67° v.a.) and 16.4 mm (1.34° v.a.) from the item center. The latter distance defines what is referred to as the outer radius or border of an item (illustrated by circles around polygons in Figure 3.2). Each individual visual item used in the stimulus displays was randomly generated from scratch to achieve maximum shape variation. Colors used for visual items were green, red, blue, yellow, black, and white.

Table 3.1: Spatial phrases used in experiment one (the same phrases were used in all other experiments). Source sets list the different candidates that filled the respective slots to form different spatial phrases.

General form	article	target item	spatial term	reference item
Example	**"Das**	**Grüne**	**links vom**	**Roten."**
Source sets	Das	Grüne Rote Blaue Gelbe Schwarze Weiße	links vom rechts vom	Grünen Roten Blauen Gelben Schwarzen Weißen
English translation				
Example	**"The**	**green [one]**	**to the left of**	**[the] red [one]."**
Source sets	The	green [one] red [one] blue [one] yellow [one] black [one] white [one]	to the left of to the right of	[the] green [one] [the] red [one] [the] blue [one] [the] yellow [one] [the] black [one] [the] white [one]

How items were combined into stimulus displays is described in the following sections. The general approach was to first arrange those items that played a specific role in the trial due to being mentioned in the spatial phrase, such that they instantiated the spatial term described by the phrase. The resulting configuration was then placed at the screen position where it would appear in the trial (the same configuration was presented at different positions in different trials). Lastly, filler items were added at random positions around the pre-specified configuration to arrive at complex item arrays more similar to real world scenes that would naturally afford the use of spatial language to point out a specific item.

Assessing spatial term fit For placing target and distractor items in relation to reference items, spatial templates were defined that described how well different spatial positions matched a given spatial term relative to a given reference item position. The spatial templates were realized based on a fit function $f(\phi, r)$, defined over angle ϕ and radius r, providing a fit value for each spatial position around a reference item located at the coordinate origin.

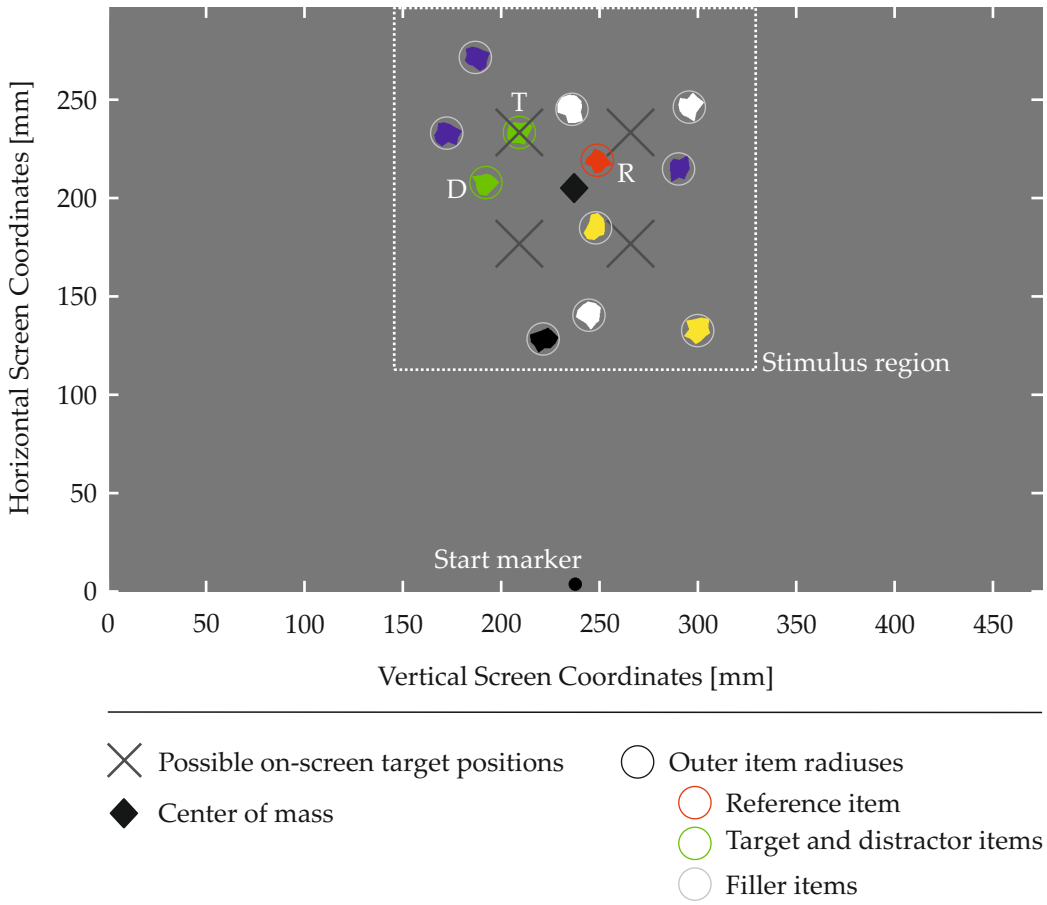


Figure 3.2: Display configuration in experiment one (experiments four, six, and seven used the same basic screen arrangement, but in part with different item configurations). The shown area spanned the entire screen. The item configuration shown in the figure was presented following the spatial phrase "The green one to the left of the red one". T denotes the target, D the distractor, and R the reference item.

In polar coordinates the function was given by

$$f(\phi, r) = e^{\left[-\frac{(\phi-\phi_0)^2}{2\sigma_\phi^2}\right]} \cdot e^{\left[-\frac{(r-r_0)^2}{2\sigma_r^2}\right]} \cdot \left(1 + e^{\beta(|\phi-\phi_0|-\phi_{\text{flex}})}\right)^{-1}, \tag{3.1}$$

where ϕ denotes polar angle, r is the radius, ϕ_0 is the mean of a Gaussian function over angle, σ_ϕ is that Gaussian function's standard deviation, r_0 and σ_r are analogous parameters for a Gaussian over radius, β is the steepness of a sigmoid function over angle, and ϕ_{flex} is the separation of its inflection point from the mean of the Gaussian over angle. The parameters used were $\sigma_\phi = 1.05$, $r_0 = 0\,\text{mm}$, $\sigma_r = 47\,\text{mm}$, $\beta = 25$, and $\phi_{\text{flex}} = 1.45$. The parameter ϕ_0 differed between spatial terms, with 'right of', 'above', 'left of', and 'below' corresponding to $\phi_0 = \{0, \frac{\pi}{2}, \pi, \frac{3}{2}\pi\}$ radians. The resulting shape of the spatial templates is shown in Figure 3.3 for each spatial term, and was inspired by behavioral data (Logan & Sadler, 1996; Hayward & Tarr, 1995; see

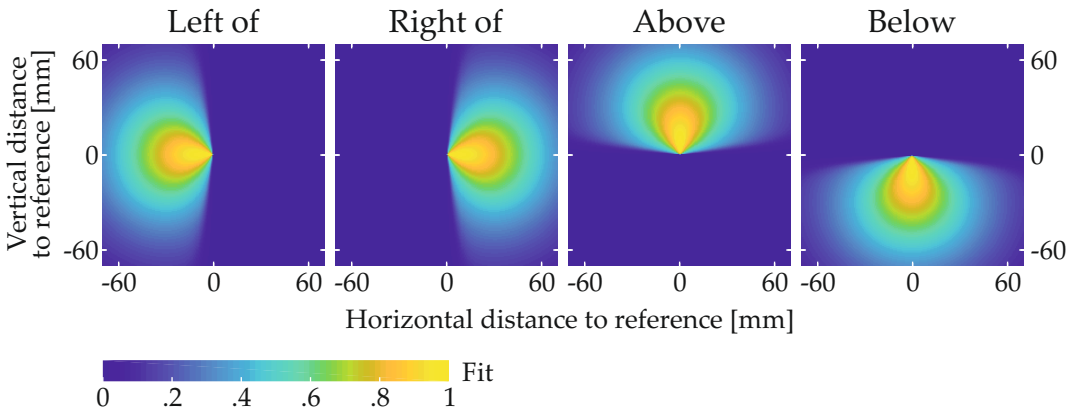


Figure 3.3: Spatial templates for the different spatial terms. These were used in the construction of visual displays for all experiments.

Section 1.7) which has been reproduced by a computational model based on similar functions (Lipinski et al., 2012).

Generating visual displays A total of 5360 different visual displays were created for the current experiment. The first step in creating these displays was to construct multiple three-item configurations, each consisting of a reference item, a target item, and a distracter item. The configurations differed in where the target item and the distracter item were placed in relation to the reference. How the configurations were generated will be described for the spatial term 'left of', but the procedure and resulting positions of the three items were identical for the other spatial terms except for rotation around the reference item.

The goal of the procedure of trial generation described in the following was twofold. First, for each spatial term, the spatial region of possible target placements (relative to the reference) was to be sampled approximately uniformly. The extent of this region was defined through a range of fit values with fixed upper and lower bounds. Second, for each target position resulting from this, the region of possible distractor placements was to be sampled approximately uniformly as well. This region as well was defined by a range of fit values. Its lower bound was fixed, as above, but its upper bound was defined in dependency of the fit value of the target position at hand, such that distractor fit was always lower than target fit.

The region generally eligible for target (center) placement included positions where $f(\phi, r) > 0.6$. As an additional constraint, the outer radiuses of reference and target item had to be separated by at least 0.5 mm (0.04° v.a.; the region where target centers could be placed is illustrated by the dotted red outline in Figure 3.4a). Target positions were centered on the junctions of

a square grid superimposed on the region resulting from these constraints, in order to achieve an approximately equal sampling of space within the region. The spacing of the grid was adjusted such that 16 possible target positions resulted (red dots in Figure 3.4c).[2]

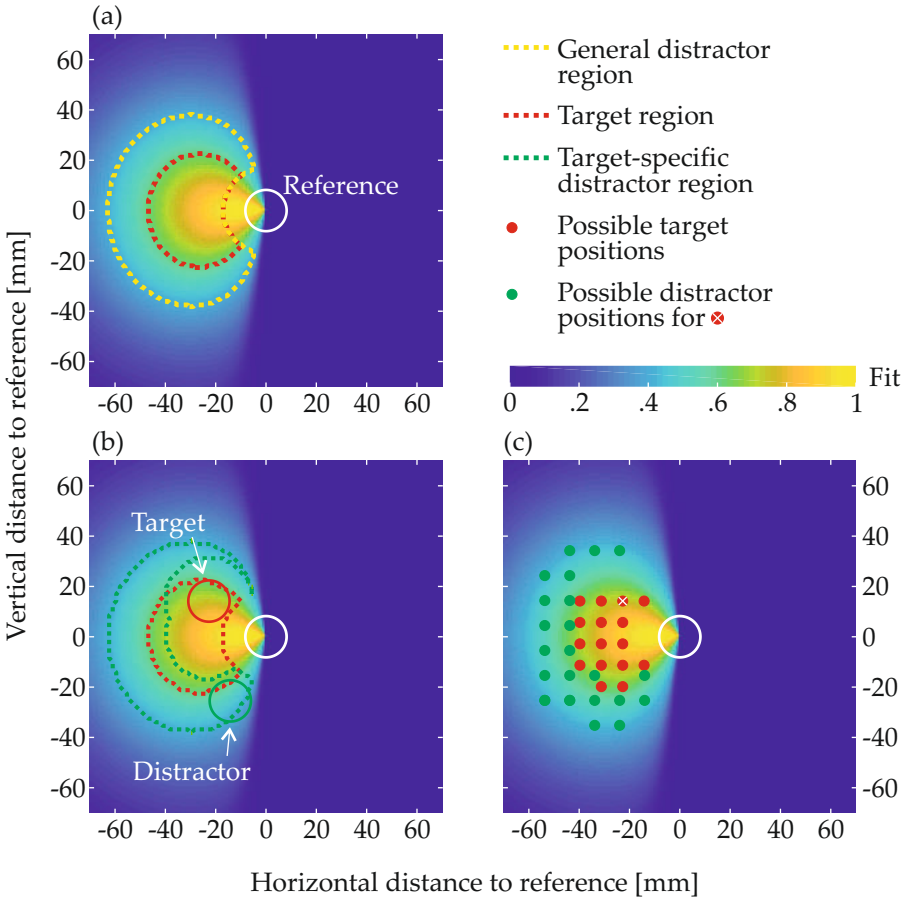


Figure 3.4: Item placement in experiment one based on fit constraints using the spatial template for 'left of'. (a) Regions generally eligible for target and distractor placement defined only by fit cut-off and minimal item distance (item centers must be placed within these). (b) Exemplary placement of target and distractor item. Note how the distractor region for this specific target placement shrinks compared to (a). (c) All possible target positions for 'left of' and all possible distractor positions for the marked target position. Large circles depict outer radiuses of the items.

For each of the 16 target positions, a separate set of distractor item positions was determined (e.g., Figure 3.4c, where green circles indicate all possible distractor positions for the target position marked with a cross). Out of these distractor positions one was used per trial (e.g., the green circle in Figure 3.4b), paired with the respective target position. The distractor positions were obtained with the same method as the target positions, but different

[2]Note that the positioning of the grid for target placement was adjusted in experiments two through five such that target positions were distributed symmetrically on either side of the main axis of the spatial term at hand, as shown in Figure 3.13.

constraints governed the shape of the regions eligible for distractor placement. First, the regions were generally constrained by $f(\phi, r) > 0.4$ (dotted yellow outline in Figure 3.4a). Second, the fit value of distractors had to be at least 0.03 lower than the fit of the corresponding target position. Third, minimum border-to-border distance between items was again 0.5 mm. Due to these constraints being defined relative to each individual target position, the shape of the distractor region was of a different shape and size for each target position (the dotted green outline in Figure 3.4b being one example). In consequence, the number of distractor positions as well differed between targets, varying from 16 to 25 (mean 20.9). The colors for each three-item set were randomly picked from the color pool, with the constraint that target and distractor had to be colored alike and the reference item had to have a different color.

For each of the four spatial terms a set of 335 different three-item configurations was obtained in this way, differing between spatial terms only in orientation and equaling a total of 1340 configurations. Each of these configurations was in turn presented in four trials, in each of which the target was placed at another one out of four different on-screen positions (gray X's in Figure 3.2). The four possible target positions were arranged in a square around the center of the stimulus region, at a distance of 28.3 mm (2.32° v.a.) horizontally and vertically. To place the target item of a configuration in a given on-screen position, the entire three-item configuration was translated such that the target item's center was placed at the desired position. Restricting target positions to a few fixed locations and instead having the stimulus array sample space around those locations alleviated the common problem of different movement metrics for different spatial locations.

Finally, nine filler items were added to each display (items in light gray circles in Figure 3.2). Each of these was colored randomly in one of the four remaining colors. Filler locations were restricted to a square stimulus region (dotted white lines in Figure 3.2), which had a side length of 184 mm (14.72° v.a.) and whose midpoint was 200.8 mm (16.01° v.a.) straight above the start marker. Fillers were placed randomly in that region, with the constraint that the center of mass (CoM) across all 12 visual items (black diamond in Figure 3.2) had to be congruent with the center of that region (with a tolerance of ± 0.8 mm in either direction for technical reasons). This means that the CoM was virtually identical across conditions and positioned in the horizontal screen center, allowing to more easily partial out a putative bias to either the CoM of all items or to the horizontal screen center. The latter was expected

because the items appeared only as soon as upward movement was detected (see *Procedure*, p.69) while the former was expected based on evidence showing averaging reaches under target uncertainty (e.g., Chapman et al., 2010; Gallivan & Chapman, 2014). Furthermore, always centering the CoM at the same position independently of where the target was situated prevented participants to infer the approximate location of the target item from the overall position of the item array. Fillers retained a border-to-border distance of at least 0.5 mm.

Finally, as an additional incentive to evaluate the spatial relation, in some trials (27%) one filler was given the same color as the target and the distracter (although technically this item is also an alternative target, the term distractor will be used to refer only to the main distractor on the side of the reference item described by the spatial term). It had to be located on the side of the reference opposed to the spatial term, and separated from the reference along the term's axis (e.g., horizontal for 'right of') by at least 28.3 mm (2.32° v.a.).

Together, the full set of visual displays was composed of 335 three-item configurations × four spatial terms × four on-screen target positions, equaling a total of 5360 trials. The trials were randomly assigned to the participants, so that each participant completed 446 trials and one completed eight more to use the entire trial set.

Analysis

The main dependent variable was the Euclidean distance of mouse trajectory data points from the direct path to the target item. This path is in the following referred to simply as the *direct path* and was computed for each trajectory individually as the straight line passing through the first data point after crossing the velocity threshold and the center of the item selected in the trial (see Figure 3.5). Note that for considerations independent of specific trajectory data the direct path was taken to start in the center of the start marker (e.g., in *Balancing the effects of potentially confounding items*, p.81).

Movement times were were defined as the duration from movement onset (first data point after crossing of the velocity threshold) to the last data point before crossing the border of the selected item. Trajectory curvature was assessed as an exclusion variable, where strongly curved trajectories were excluded from analyses as described in *Assessing trajectory curvature* (p.79; only initial tests for bimodality of curvature value distributions included maximum curvature values from all trajectories). Accuracy was assessed in terms of percent correct responses, where responses were considered correct if par-

ticipants selected the item that was the best fit for the spatial term at hand according to the fit function (see *Assessing spatial term fit*, p.73). Incorrect trials were excluded from all other analyses.

Trajectory preparation In accord with the above definition of movement time, trajectories were trimmed to start with the first data point after crossing of the velocity threshold and to end with the last data point before crossing the outer radius of the selected item. This and the following steps of trajectory preparation are illustrated in figure Figure 3.5. In the first panel, discarded trajectory portions are marked in gray (before point a, after point b). After trimming, each trajectory was translated such that the first data point was moved to the coordinate origin, and then rotated around that point by the angle between the positive y-axis (vector $[0, 1]$) and the vector specifying the center position of the selected item (third and fourth panel in Figure 3.5). Note that this entails that final trajectory data points tended to lie not at $x = 0$ but somewhat lateral to the y-axis, depending on where the radial border of the selected item had been crossed. This ensured that any deviation affecting trajectories until the end of the movement was retained in the rotated versions. Following these transformations, positive values of deviation denote deviation to the right side of the direct path while negative values denote deviation to the left side (sides are given relative to the "direction of travel" toward the target).

Finally, each trajectory's x- and y-coordinates were linearly interpolated over 151 equally spaced steps of movement time to enable averaging. Averaging trajectories will thus combine spatial data from the same proportion of movement time elapsed since movement onset.

This time-normalization was chosen over matching data points by a shared spatial metric because there were no specific hypotheses in how far the spatial position of the mouse cursor relative to the items of interest at each time point would be linked to the strength of putative biases toward or away from these items. Hypotheses rather pertained to the impact of sequential (and thus temporally ordered) item selection by visual attention on the biases. A shared spatial metric was also precluded by the two different target distances and because trajectories could extend beyond the target item before returning to it (although this was rarely the case in the analyzed data).

Assessing trajectory curvature The degree of curvature along each trajectory was computed using the following algorithm. Three steps of the algo-

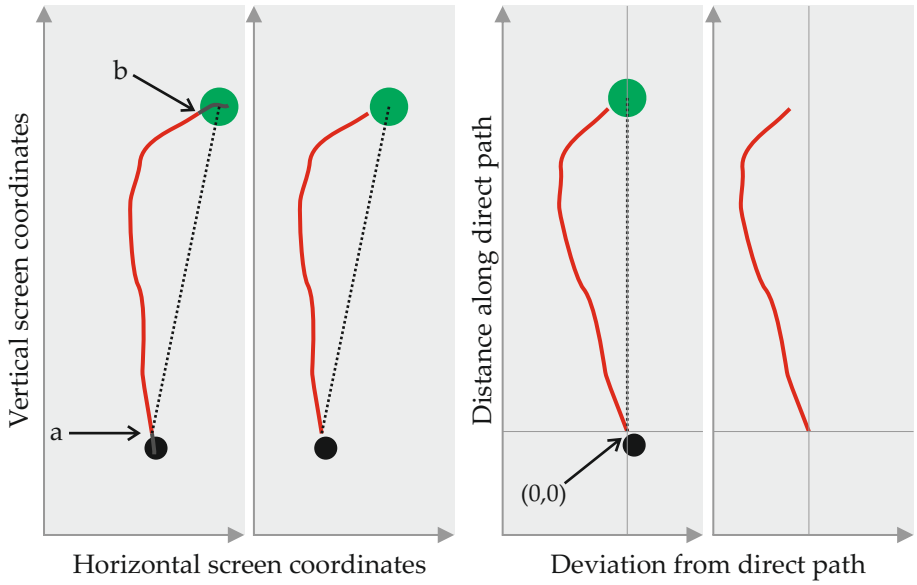


Figure 3.5: Trajectory preparation included removing the trajectory portions marked in gray in the leftmost panel, namely those preceding the first data point after velocity threshold was crossed (a) and those following the data point just before the crossing of the item border (b, target item in green). The direct path to the target (dotted line) is defined by point a and the item center. The black dot denotes the start marker. The leftmost panel shows hypothetical raw trajectory data (red), of which the second panel shows the trimmed version, the third panel shows the rotated and translated version, and the rightmost panel shows the same data without items.

rithm are illustrated in Figure 3.6. Given a raw (i.e., non-interpolated) trajectory, a fragment made up of two linear segments of 15 mm length each (red lines in Figure 3.6), linked at a central vertex, was fitted to the trajectory at different positions along its arc length, by adjusting the angle between the two segments such that all three vertices lay on the trajectory. Curvature for the trajectory arc length at the fragment's central vertex was computed as the absolute angle by which the fragment's second segment deviated from the direction of the first segment (gray arcs in Figure 3.6). This procedure was performed along the entire trajectory, for successive points separated by one millimeter steps. The maximum curvature for a given trajectory was recorded as the highest angle computed anywhere along its length. Possible curvature values ranged from zero (straight line) to π radians (antiparallel segments). Trajectories whose maximum curvature exceeded 0.933 radians were excluded from analyses.

The described method allowed to obtain curvature values independently of the varying Euclidean length of trajectory segments in the raw data. Adjusting the parameter of segment length of the fitted fragment allowed to vary the scale of trajectory turns that contributed to curvature. For instance, a larger segment length decreases the impact of small deviations from the

principal direction within a given trajectory portion and at the same time assigns higher curvature to turns formed by many short trajectory segments that change direction only gradually.

The employed segment length and the chosen exclusion threshold were determined based on a procedure of visually inspecting exclusion outcome for different segment lengths and thresholds. The parameter values were adjusted to balance two competing requirements. On the one hand, trajectories were excluded that exhibited sharp turns which changed overall trajectory direction, as may be expected when a target choice is revised abruptly. This was done to exclude trials where participants may have made an early target decision independent of the grounding process and later corrected it. It also served to exclude outliers in terms of extreme deviation that may have arisen from gross coordination failures in handling the computer mouse followed by a correction of movement direction. On the other hand, trajectories with very brief deviations that appeared to result from slightly overshooting the target or from small imprecisions in mouse handling were not excluded, in order to retain as much data as possible. Examples for discarded and retained trajectories will be shown along with the other experimental results (Figure 3.10).

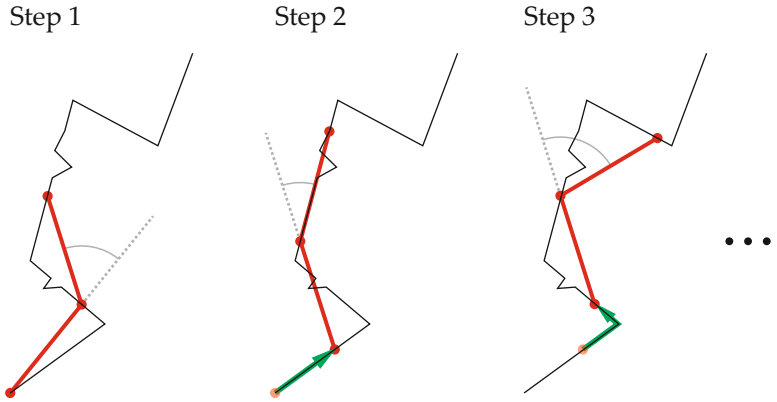


Figure 3.6: Schematic illustration of three steps of the curvature computation algorithm. Black lines show part of a hypothetical raw trajectory. Red lines illustrate the two-segment fragment fitted to a different portion of the trajectory in each step. Green arrows denote the fixed shift distance of the fragment's first vertex along the trajectory arc length from step to step. Angles computed in each step are shown as gray arcs.

Balancing the effects of potentially confounding items The current paradigm did not allow to manipulate the independent variables, that is, the side of different items relative to the direct path, in full independence of each other. This was because the spatial placement of the different items was constrained by the position of the other items in the same trial, to satisfy spatial terms and fit constraints (see *Generating visual displays*, p.75). The approach used here was

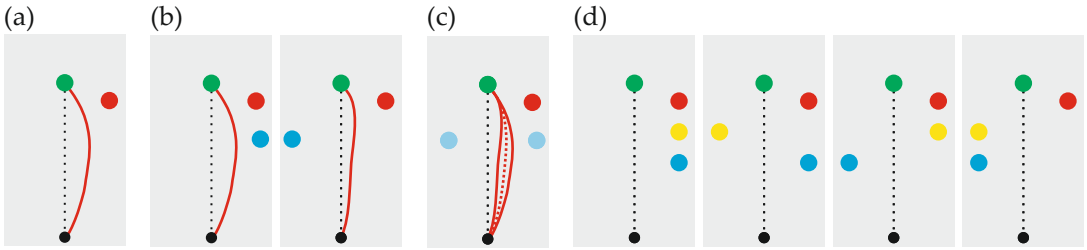


Figure 3.7: Schematic depiction how the "pure" impact of effect sources was isolated through balancing item sides across averaged trials. The green item is the target, solid red lines are hypothetical trajectories, the direct path is indicated by a dotted black line. In (a), the deviation from the direct path is clearly attributable to the red item of interest, while in (b) the deviation cannot be attributed to the red or blue item. (c) Averaging trajectories over panels in (b) results in a mean trajectory (dotted red line) that allows to judge the deviation caused by the red item (assuming no interactions). (d) Instances of the four balancing categories that arise with two confounding items (blue and yellow).

to first generate a set of trials as described above, and only later assign them to different *balancing categories* based on the relationship of the item locations, such that counterbalanced means arose which could be subjected to statistical testing (the balancing categories are, with some restrictions, analogous to the cells of statistical procedures such as ANOVA).

As described initially, the goal of the experiment was to examine whether trajectories were attracted or repelled by visual items situated on the left or right side of the direct path to the target. The rationale behind the statistical tests described in the next section (*Statistical analysis*, p.90) was to test the effect of a given item of interest, such as the distractor, by comparing average trajectories over trials in which this item of interest was to the left of the direct path to those where it was to the right of that path. Such a comparison and its interpretation would be straightforward for visual displays containing only a target item and the item of interest (Figure 3.7a) since in the absence of confounds an observed difference could be reliably attributed to the latter.

In the current experiment, however, both the distractor and the reference item were necessarily present in each display, and every trajectory was potentially affected by both of their positions. To examine the individual effects of distractor and reference, separate statistical comparisons were conducted, and in each of these comparisons one was the item of interest while the other posed a potential confound. For this more complex case with one item of interest and one confounding item, there are two principal item arrangements or *balancing categories*: Either the item of interest and the confounding item are on the same side of the direct path, or they are on different sides (Figure 3.7b).

Assuming no interaction effects, the effect of the confound can be removed by averaging over the two balancing categories (Figure 3.7c), with the im-

portant prerequisite that the data from each balancing category be weighted equally strong. This principal approach was applied throughout analyses here. Equal impact of the data from each balancing category was ensured by first computing across-trial means within individual balancing categories and then averaging over these means. This two-step averaging was used instead of averaging over trials directly because the experimental design did not allow to obtain equal trial numbers for each balancing category.[3]

In the current experiment, however, there was a third confounding factor present in each display, namely the CoM. As described in *Generating visual displays (p.75)*, the CoM was congruent with the horizontal screen center by design and the biases based on these two were expected to be congruent in directionality and to be superposed with the effects of distractor and reference. Thus, there were three distinct potential effect sources in each trial. In a given comparison focusing on the effect of one of them, the remaining two sources played the role of potential confounds. Note that for simplicity the term 'items' is used in the remaining description of balancing to refer to the distractor, the reference, and the CoM although the latter is not a single visual item.

Two confounding items result in four balancing categories which are illustrated in a schematic manner in Figure 3.7d. Balancing was still ensured as described above, by combining means from all four balancing categories into overall means. Labels for the four balancing categories underlying each comparison are based on whether the two confounding items were on the same side of the direct path as the item of interest (label code *s*) or on a different side (label code *d*), and the three relevant items are denoted by *d* for distractor, *r* for reference, and *c* for the CoM. For instance, the label *rs/cd* denotes the balancing category in which the reference item was on the same side of the direct path as the distractor, and the CoM was on the opposite side. Table 3.2 lists the balancing category labels relevant for the distractor effect (left part) and for the reference effect (right part).

An additional complication for balancing in the current experiment was

[3]This had several reasons. One was that some trials were lost by being discarded due to exceeding curvature threshold or because of incorrect responses, neither of which could be controlled in advance. Also, the balancing category of a trial could not be predicted perfectly in advance, as the direct path and thus the side of the items depended in part on where in space the velocity threshold was crossed (although this was a marginal influence). Most importantly, the basic trial set did not include equal trial numbers for each balancing category to start with (assessing the direct path preliminarily as beginning in the center of the start marker), as some categories were more likely to be generated by the algorithm used to construct visual displays (e.g., due to differing sizes of the spatial areas in which items had to be placed to result in a specific balancing category).

Table 3.2: Balancing category labels for the distractor effect (left table) and the reference effect (right table). 'Side' refers to the side of the respective item relative to the direct path in comparison to the side of the item of interest relative to the direct path.

	CoM side	
Reference side	Same	Different
Same	rs/cs	rs/cd
Different	rd/cs	rd/cd

	CoM side	
Distractor side	Same	Different
Same	ds/cs	ds/cd
Different	dd/cs	dd/cd

that which side of the direct path the confounding items could be placed on was strongly constrained in some conditions (i.e., certain combinations of spatial term and on-screen target position).

First, in trials using the horizontal axis spatial terms, that is, 'left of' or 'right of', the side of the reference item relative to the direct path had a fixed coupling to the spatial term. In trials using the spatial term 'left of' the reference necessarily was on the right side, and in trials using the spatial term 'right of' it was on the left side. This coupling resulted from the fact that the visual relational pairs were constructed to match the spatial term in the trial by placing the target within a region constrained by a relatively strict fit value cut-off (see *Generating visual displays*, p.75). Note that this problem was not practically relevant for the vertical axis spatial terms ('above' and 'below'), because the horizontal position of the reference item could vary more freely in these trials, allowing the reference to be placed on either side of the direct path without violating fit constraints.

Second, the side of the CoM was prescribed by on-screen target position. When one of the two positions to the left of the screen center was used, the CoM was on the right side of the direct path, and when one of the two right positions were used, the CoM was on the left side of the direct path. This resulted from the fact that both the possible on-screen target locations and the CoM location were fixed.

As a consequence of these fixed couplings it was also fixed for each condition whether or not it allowed item configurations where the reference and the CoM were on the same or on different sides of the direct path. In other words, certain balancing categories could not be realized in some conditions. To illustrate this problem, Table 3.3 shows exemplary item configurations for all combinations of spatial term and on-screen target position, and indicates for each of these conditions which of the four balancing categories for the distractor effect were possible or impossible to realize. For instance, for the spatial term 'left of' in conjunction with the upper left target position (top left

panel in Table 3.3), it was not possible to realize rs/cd and rd/cs, because the CoM and the reference were both on the right side of the direct path so that either both or neither of them could share a side with the distractor.

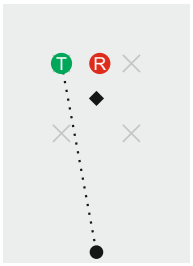
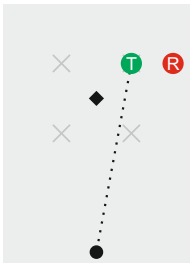
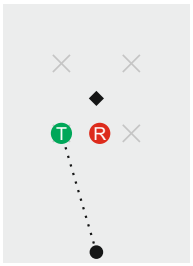
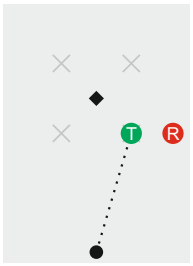
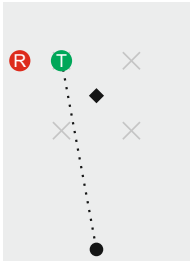
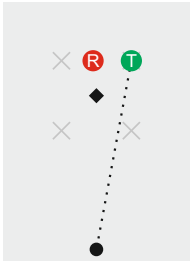
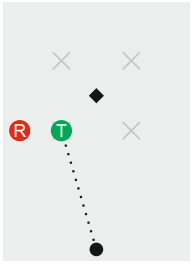
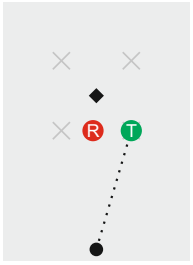
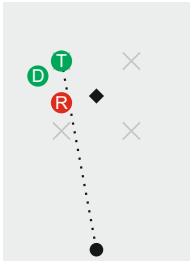
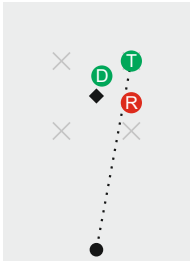
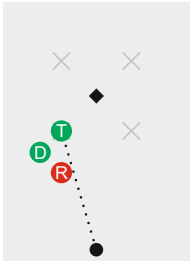
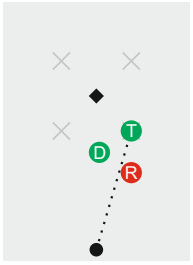
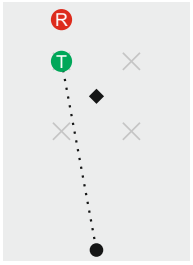
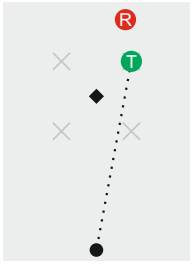
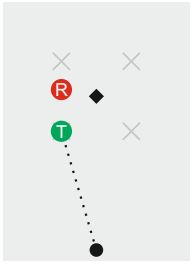
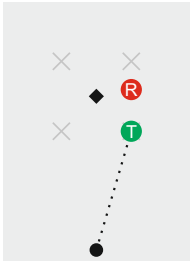
Due to these inevitable imbalances it was not possible to simply compute a balanced mean over the four balancing categories for each individual condition. Instead, sets of conditions were combined such that overall all four balancing categories were represented in the combination. This was done by averaging over all balancing category means available in the combined conditions. Statistical comparisons were then only conducted between the resulting balanced overall means.

Which conditions were combined in this manner is illustrated in Figure 3.8 for comparisons examining the distractor effect and in Figure 3.9 for comparisons examining the reference effect. These figures list all conditions and the balancing categories that were realized in each of them, illustrating the associated item configurations by schematic depictions of the respective visual displays. In the figures, color and style of panel outlines indicate which conditions were combined into overall means. Panels sharing both outline color and style were averaged, while comparisons were conducted between the overall means of sets sharing the same color. In addition, overall comparisons were performed between means over the two larger sets of differing outline style (Concrete statistical tests for the different comparisons are described in detail in the following section). Note that for all comparisons each combined set includes an equal number of means from each balancing category. Note furthermore that the data illustrated in the two figures are identical but arranged differently to better show the differing combinations and comparisons performed for the two effects.

Figure 3.8 and Figure 3.9 also show that the estimates obtained for each effect were based on partly differing sets of conditions. Consider, for instance, Figure 3.8 and the comparison of *distractor left* to *distractor right* for horizontal axis spatial terms (dotted blue outlines versus solid blue outlines). Even though each of the compared overall means included two instances of each balancing category, they differ in that corresponding balancing categories come from different combinations of spatial term and on-screen target position. Thus, a necessary assumption that was made to allow interpreting the obtained effect estimates is that there were no pronounced interaction effects between spatial term, on-screen target position, and the side of the item of interest.

An important special case where this assumption does not hold is posed

Table 3.3: Overview of which balancing categories can and cannot be realized for the different combinations of spatial term and on-screen target position (*y* for yes, *n* for no) by placing the distractor item. The second and third row denote the balancing category for each column depending on the side of reference and CoM, where *s* (same) and *d* (different) denote whether the respective item shares the distractor side. Green and red dots denote target (T) and reference (R), respectively, dotted lines represent the direct path to the target, the center of mass across all items (fillers not shown) is depicted by a black diamond, the start marker is shown as a black dot, and light gray crosses indicate possible on-screen target positions.

Target Pos.	upper left				upper right				bottom left				bottom right			
Reference	s	s	d	d	s	s	d	d	s	s	d	d	s	s	d	d
CoM	s	d	s	d	s	d	s	d	s	d	s	d	s	d	s	d
Spt.: Left																
Possible?	y	n	n	y	n	y	y	n	y	n	n	y	n	y	y	n
Spt.: Right																
Possible?	n	y	y	n	y	n	n	y	n	y	y	n	y	n	n	y
Spt.: Above																
Possible?	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
Spt.: Below																
Possible?	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y

by reference side in the case of horizontal spatial terms (Figure 3.9, dotted blue outlines versus solid blue outlines). There, reference side and spatial term are coupled inseparably: The reference item is always on the left side of the direct path when the spatial term is 'right of' and vice versa. When interpreting the two comparisons that include these sets of conditions it must be kept in mind that they show the combined effects of reference side and spatial terms (i.e., *reference left* and 'right of' versus *reference right* and 'left of'),

Note that balancing took into account only the side of items, whereas the Euclidean distance of items from the direct path was not strictly controlled. Finally, note that the above balancing method was applied not only to trajectories, but also to other trial-specific data such as movement times (and when reporting overall data not affected by balancing, this is explicitly noted).

Figure 3.8 (on following page): Schematic depiction of possible item configurations in experiment one and four for each combination of distractor side, on-screen target position, and spatial term. Balancing categories are indicated in the bottom right of each panel. Panels are arranged by distractor side to illustrate which conditions were combined and compared to examine the distractor effect. Means from panels sharing outline color and style were averaged, and the resulting means were compared between sets of differing outline style (within and across colors). Target (T) and distractor (D) are shown as green dots, the reference (R) is shown in red, and the center of mass across all items (fillers not shown) is depicted by a black diamond. The start marker is shown as a black dot and light gray crosses indicate possible on-screen target positions.

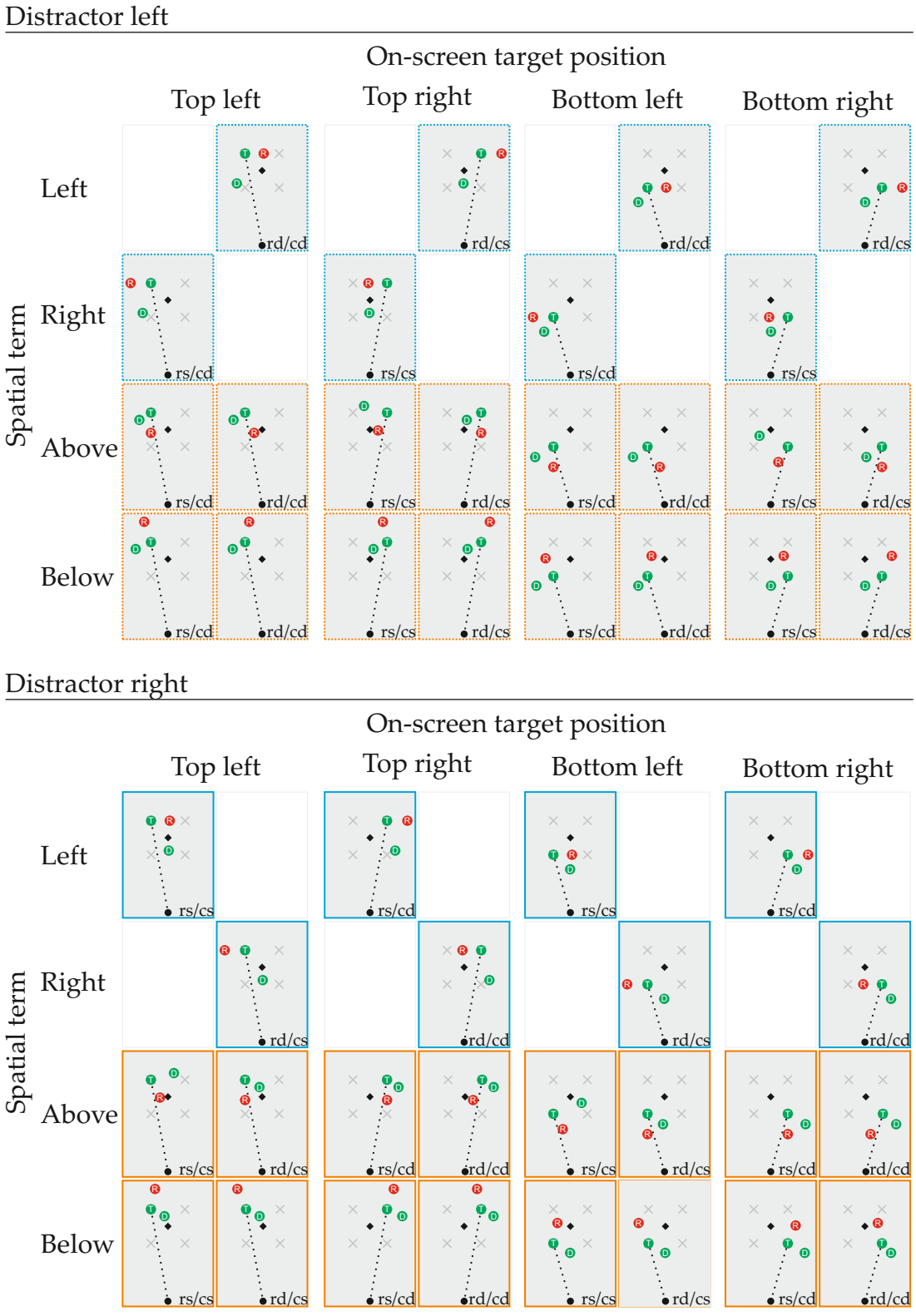


Figure 3.8: Caption on previous page.

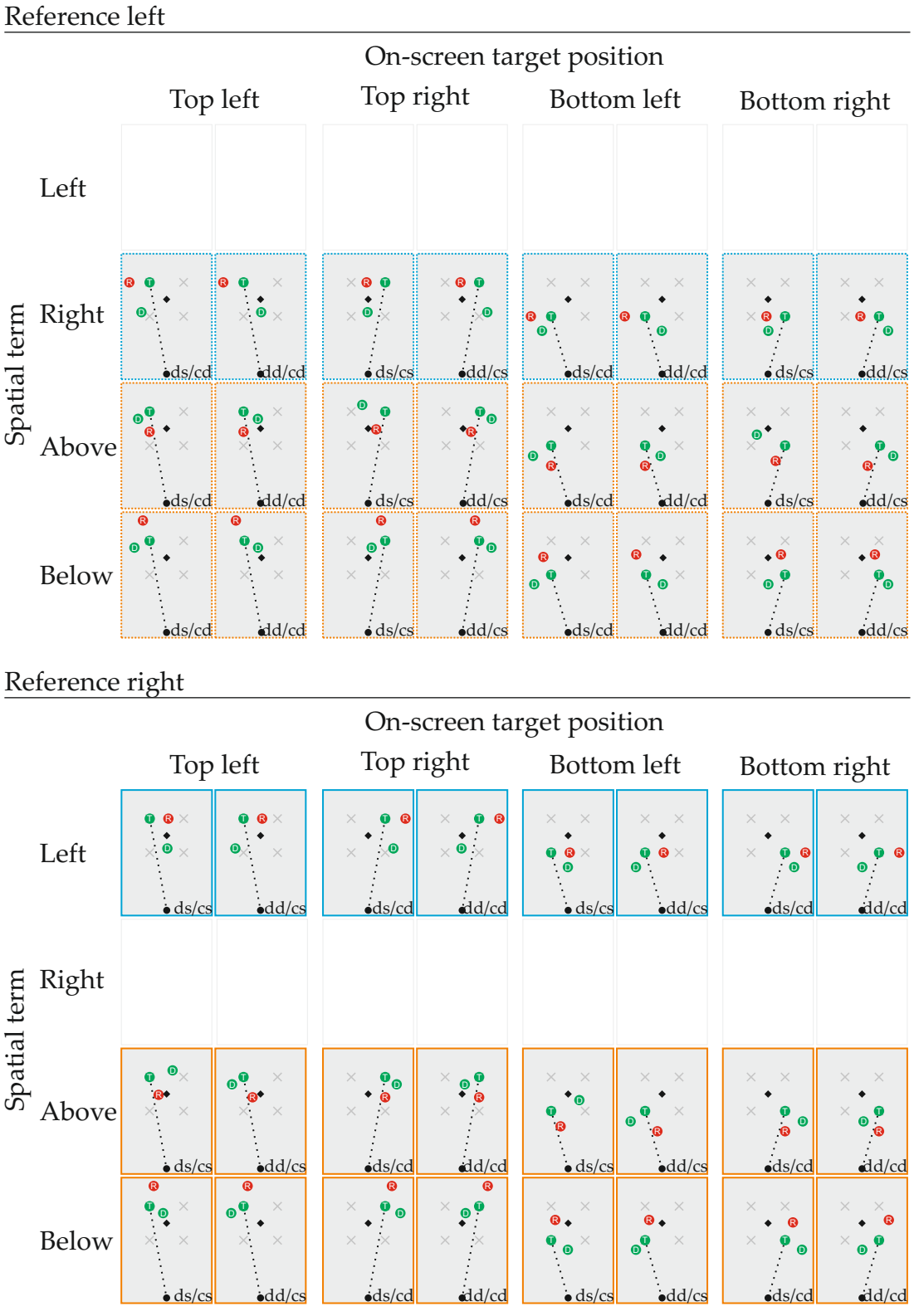


Figure 3.9: Caption on following page.

Figure 3.9 (on previous page): Schematic depiction of possible item configurations in experiment one and four for each combination of reference side, on-screen target position, and spatial term. Balancing categories are indicated in the bottom right of each panel. Panels are arranged by reference side to illustrate which conditions were combined and compared to examine the reference effect. Means from panels sharing outline color and style were averaged, and the resulting means were compared between sets of differing outline style (within and across colors). Target (T) and distractor (D) are shown as green dots, the reference (R) is shown in red, and the center of mass across all items (fillers not shown) is depicted by a black diamond. The start marker is shown as a black dot and light gray crosses indicate possible on-screen target positions.

Statistical analysis Trajectory data were subjected to repeated measures analyses as described in the following. The balanced composition of means subjected to the statistical tests has been described in the preceding section (*Balancing the effects of potentially confounding items*, p.81). Within-subject factors for analysis included *reference side*, with the levels *reference left* and *reference right* (of the direct path), *distractor side*, with the levels *distractor left* and *distractor right* (of the direct path), and *spatial term axis* with the levels *horizontal* and *vertical*.

To assess the effect of reference side on trajectories, three planned contrasts compared mean trajectories between left and right conditions. One compared trajectories across spatial terms, one included only horizontal-axis spatial terms, and one included only vertical-axis spatial terms. The effect of distractor side was assessed with three analogous contrasts. Each planned contrast consisted of 151 paired sample t-tests with $p < .01$, comparing mouse x-coordinates (i.e., deviation from the direct path) at each time step. Reference side and distractor side were not analyzed in a single factorial ANOVA because there were no hypotheses as to an interaction between them.

In sum, a total of six planned comparisons between mean trajectories were conducted on the data from this experiment. Inflated Type I error risk was addressed by choosing $p < .01$ for each t-test. Although numerically this does not fully account for the six comparisons, it was deemed sufficient given that contrasts were preplanned and hypothesis-guided (Armstrong, 2014), and due to the non-independence of data sets underlying some of the contrasts (comparisons by spatial term axis and across spatial terms), which tends to increase the conservativeness resulting from correction procedures (Winer et al., 1991; Abdi, 2007).

As an additional indicator for the overall significance of trajectory differences that accounts for the multiple tests conducted for each comparison, a bootstrap procedure was performed as described in the next section. The method provides a length criterion for the number of sequentially significant

t-tests required for overall significance (Dale et al., 2007). The overall length criterion was as well based on $p < .01$.

Comparisons between trajectories were complemented by effect size calculations at each time step using Cohen's d_z, a variant of Cohen's d for paired samples (namely the mean over difference scores divided by their standard deviation; Cohen, 1988; Lakens, 2013). Movement time means were computed analogous to mean trajectories. They were not statistically compared as there were no hypotheses in this regard.

Finally, the distribution of maximum curvature values was examined for signs of bimodality. Bimodality was assessed over all correct trajectories, including those exceeding the curvature threshold and, if bimodality was observed in this sample, also for the smaller set of trajectories which resulted from the exclusion of the trajectories exceeding the curvature threshold. This was done to determine, first, whether two distinct populations of trials were at all discernible (see Section 1.6) and, second, whether trials from both of these populations may still have affected the ultimately analyzed set of trajectories.

Bimodality was assessed using Hartigan's dip test (J. A. Hartigan & Hartigan, 1985; P. M. Hartigan, 1985) in the MATLAB implementation by Mechler (2002). It tests the null hypothesis of unimodality against the alternative hypothesis of multimodality, with p values below .05 indicating bimodality. While the bimodality coefficient (SAS Institute, 2012) has been more widely used for detecting bimodality in response distributions as a sign for distinct processing modes (Freeman & Dale, 2013), it is prone to detecting bimodality erroneously when skewness is high (Pfister et al., 2013). Since the distribution of maximum curvature values obtained here was positively skewed, the more robust dip test was used instead.

Bootstrapping As described above, mean trajectories were compared by means of pairwise statistical tests at each time step, resulting in a multitude of tests. The question arises how many successively significant comparisons are required to safely assume that a given trajectory portion is indeed overall significantly different.

If successive tests were independent of each other, this could be answered by considering the sequence of tests as a Bernoulli process with success probability equal to the alpha level used for each test and determining the longest success run length whose probability of occurrence is just below the desired

overall alpha level.[4] However, successive data points in the mean trajectories are highly interdependent (Dale et al., 2007) due to the physical constraints governing hand and mouse movement, rendering this reasoning inadequate.

A recent perspective on the issue is to consider sequences of statistical tests over trajectory differences as units that stand for a single comparison of the trajectories as a whole, thus rejecting the need for alpha correction as long as the outcome of the comparison is presented and interpreted in its entirety (Gallivan & Chapman, 2014; Chapman, 2011). In other words, exactly due to the strong physical interdependence of consecutive data points in natural movement trajectories, the problem of multiple comparisons is attenuated.

However, many researchers in mouse tracking have adopted a bootstrap approach (Efron & Tibshirani, 1993) that was first introduced by Dale et al. (2007) and has since been applied in many studies (e.g., Bartolotti & Marian, 2012; Anderson et al., 2013; Duran et al., 2010; Freeman et al., 2008; Scherbaum et al., 2015). The method preserves the dependency between time steps and yields an empirical distribution of bootstrap replications over the maximum length of significant sequences. Based on a pre-specified p value, a criterion for sequence length in the real data is derived beyond which the presence of an overall effect is assumed. In keeping with the mouse tracking literature, this approach was adopted here as an additional indicator for overall significant trajectory divergence.

The method was implemented according to the description provided by Dale et al. (2007; see also, Scherbaum et al., 2015). For each reported comparison, a separate criterion was computed, using the compared data as basis. Computing a criterion for a given comparison included 10,000 bootstrap replications of maximum sequence length, each obtained as follows.

Given two overall mean trajectories computed from real experimental data, one per condition, each was used to construct one artificial mean trajectory for each participant. Artificial trajectories were constructed by drawing for each time step from a normal distribution that was defined by the mean and standard deviation at that time step in the condition at hand. The artificial trajectories were then subjected to statistical testing in the same manner as the real data, obtaining as many test results as time steps.

[4]Determining this length threshold via 100,000 simulations of a Bernoulli process with a success probability of one percent (the alpha level used for most comparisons here) and a length of 151 (the number of tests for each comparison reported here) showed that three successes in sequence occurred in less than one percent of all repetitions. This may serve as a liberal criterion.

For each of the 10,000 repetitions the length of the longest consecutive sequence of significantly different time steps was recorded. For a given comparison, the final criterion for overall significance was equal to the number of successively significant steps in the longest sequence that occurred in less than $p * N_{bootstrap}$ artificial experiments, where p is the desired overall alpha level for the comparison at hand, and $N_{bootstrap} = 10,000$ is the number of artificial experiments performed. Bootstrapping for ANOVAs (experiment six) worked in the same way, but with more than two involved conditions and thus more than two artificial trajectories per participant. In this case, a separate criterion was obtained for each main effect and interaction.

A drawback of this canonical procedure is that the criterion becomes increasingly conservative as spatial separation between mean trajectories increases relative to variance. This is because very divergent trajectory portions that differ significantly in the real data tend to remain significantly different in the bootstrap replications, since re-sampling is unlikely to remove significance in this case (while for smaller significant effects it is more likely to be removed). The result are longer sequences of significance in most bootstrap replications and therefore stricter overall length criteria. Due to this, the obtained threshold may become overly conservative for trajectories that differ markedly over long stretches (as opposed to the situation where the difference between data points is small compared to variance, which has been the usual case in most mouse tracking studies). Missing overall significance must therefore be interpreted with care in the case of trajectory divergences with large effect sizes.

Note that other authors (Freeman et al., 2008; Duran et al., 2010; Anderson et al., 2013; but see Bartolotti & Marian, 2012; Scherbaum et al., 2015) have used a fixed criterion of eight time steps as their criterion for overall significance (for overall $p < .01$ and $p < .05$ for each of 101 successive tests), based on the value originally determined by Dale et al. (2007). However, as the previous paragraph illustrates, and as is generally the case in bootstrap scenarios, the obtained criterion is dependent on the source data on which the artificial experiment is based, so that it is more reliable to compute a separate criterion for each comparison based on the experimental data at hand (especially when effect sizes strongly differ between comparisons).

3.1.2 Results

When asked, participants reported not to have noticed that possible target positions were restricted to four on-screen locations (some noted that targets tended to be located around the center area of the item arrays rather than in the outer regions). Movement onset was generally registered close to the center of the start marker (M = 2.14 mm, SD = 1.97 mm).

A total of 5245 trajectories was obtained (115 were lost due to technical issues). Of these, 5003 (95.39 %) were below curvature threshold (M = 416.92, SD = 35.91 equaling M = 95.3 %, SD = 2.76 %). Of the non-curved trajectories, 90.17 percent (4511) were correct responses and thus entered further analysis (86.01 % of all obtained trajectories).

Participants achieved a mean accuracy of 90.18 percent (SD = 3.34 %) and their mean movement time was 1073 milliseconds (SD = 112 ms). Note that the numbers reported so far were not affected by balancing but based on simple averaging over the respective trial ensembles.

Mean data reported from here on is based on balancing as described in Section 3.1.1. Condition-specific movement times are listed in Table 3.4 and showed only marginal differences.

Table 3.4: Movement times and standard deviations (SD) for experiment one.

	Distractor side				Reference side			
	Left		Right		Left		Right	
Spatial terms	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Overall	1072	113	1075	113	1074	115	1073	110
Left/Right	1088	104	1085	114	1092	110	1081	108
Above/Below	1065	123	1067	111	1060	119	1072	117

Figure 3.10 shows fifty examples for trajectories below and above curvature threshold (panels a and b, respectively) along with the empirical distribution over maximum curvature value for all correct responses (panel c), with red bars indicating curvature above threshold (i.e., trajectories excluded from other analyses). For the distribution Hartigan's dip test indicated no bimodality ($p > .05$).

The top row of panels in Figure 3.11 visualizes the results of comparisons of mean trajectories by distractor side, where red and blue circles labeled 'D' in the top of each panel indicate distractor side for the correspondingly colored mean trajectory.

Across spatial terms, trajectories diverged in a way consistent with a bias

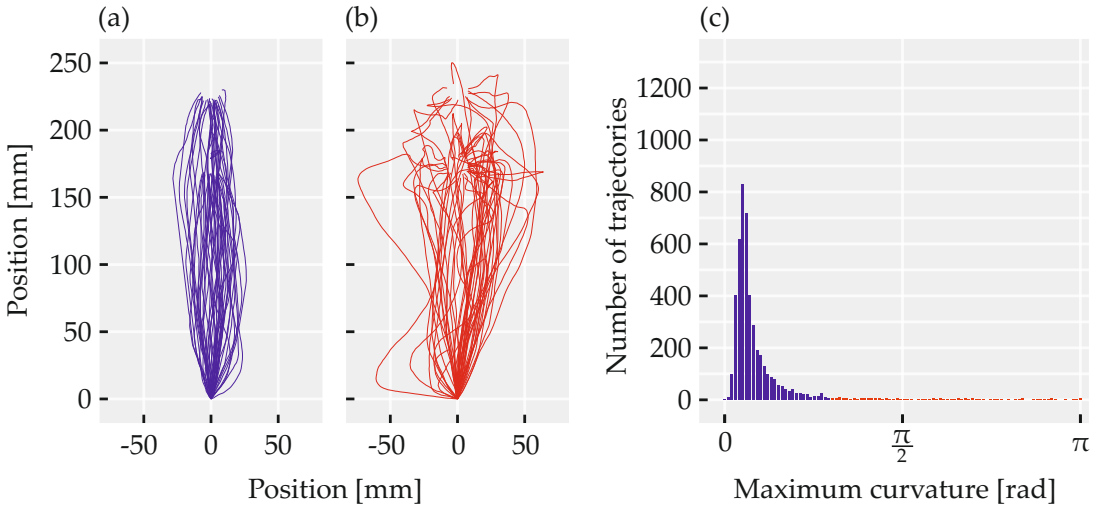


Figure 3.10: Fifty randomly selected trajectories obtained in experiment one that were (a) below curvature threshold and (b) exceeding curvature threshold. Panel c shows the overall distribution of trajectories over maximum curvature values, where red bars correspond to trajectories that were discarded due to high curvature. Only correct responses are shown.

toward the distractor (Figure 3.11a), with 106 successive time steps showing significant differences at $p < .01$, thus exceeding the bootstrap criterion ($p < .01$) of 18 time steps. The sequence of significant differences extended from 30.46 to 100 percent of movement time, with the minimum p value occurring at 94.7 percent movement time ($t(11) = 11.74$, $p < .001$, $d_z = 3.39$).

For horizontal axis spatial terms, the bias toward the distractor was present as well (Figure 3.11b), with 85 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 8 time steps. The sequence of significant differences extended from 44.37 to 100 percent of movement time, with the minimum p value occurring at 75.5 percent movement time ($t(11) = 7.72$, $p < .001$, $d_z = 2.23$).

Similarly, for vertical axis spatial terms, the bias toward the distractor was present (Figure 3.11c), with 92 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 14 time steps. The sequence of significant differences extended from 39.74 to 100 percent of movement time, with the minimum p value occurring at 98.01 percent movement time ($t(11) = 6.85$, $p < .001$, $d_z = 1.98$).

The bottom row of panels in Figure 3.11 visualizes the results of the comparisons by reference side, where red and blue circles labeled 'R' in the top of each panel indicate reference side for the correspondingly colored mean trajectory.

Across spatial terms, a mixture of two biases was visible (Figure 3.11d). In the first half of movement time, trajectories diverged in a way consistent with

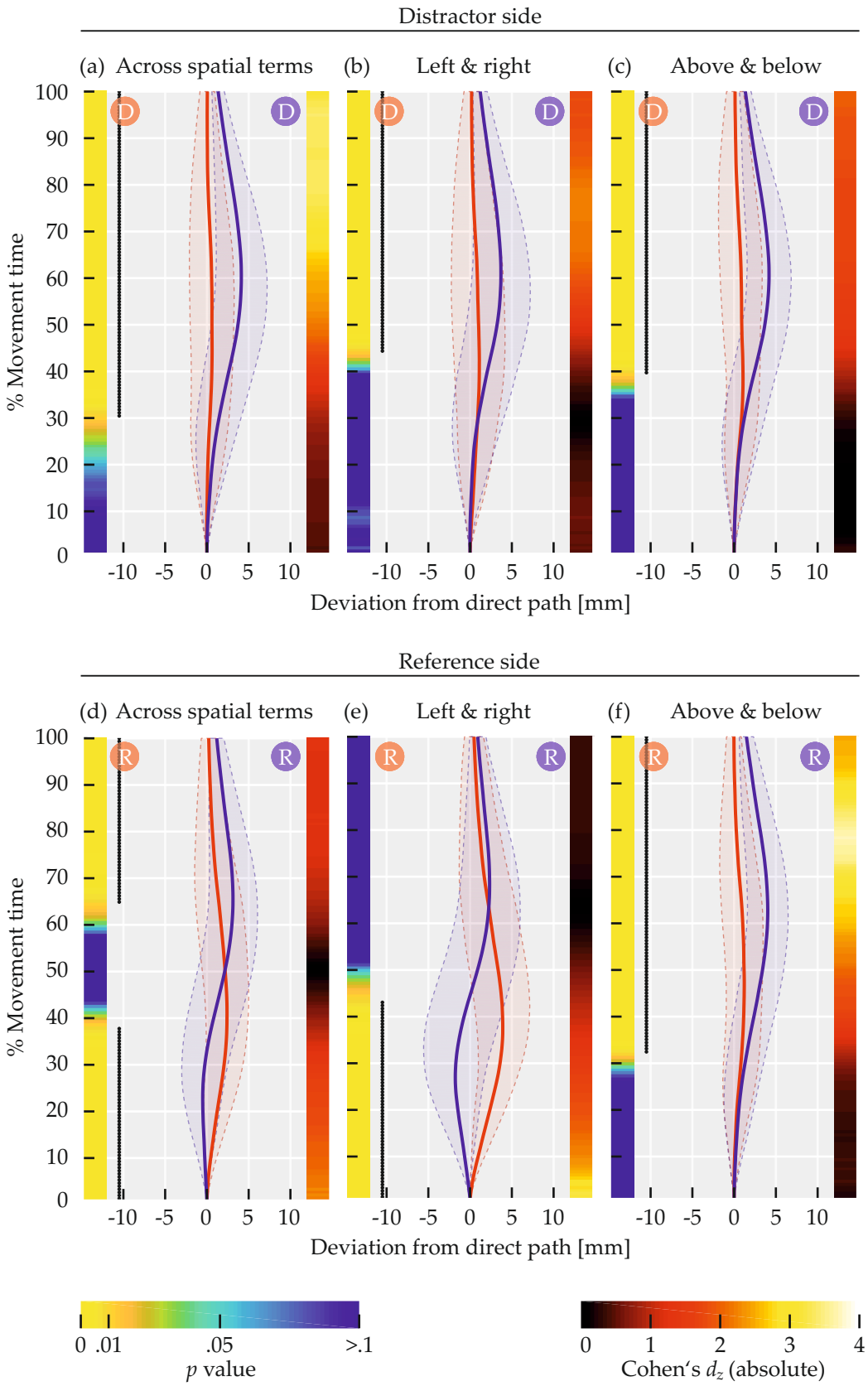


Figure 3.11: Comparisons of mean trajectories for experiment one. Red and blue circles labeled 'D' or 'R' in the top of the panels indicate distractor or reference side for the correspondingly colored mean trajectory. Transparent regions delimited by dashed lines indicate between-subjects standard deviation. Left image maps in panels indicate p values from t-tests at that time step, right image maps indicate effect sizes (absolute Cohen's d_z). Black dotted lines on the left span time steps where mean trajectories differed significantly.

a bias away from the reference item. This effect spanned 56 successive time steps with significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 6 time steps. For this effect, the sequence of significant differences extended from 1.32 to 37.75 percent of movement time, with the minimum p value occurring at 1.99 percent movement time ($t(11) = -7.93$, $p < .001$, $d_z = -2.29$).

In the second half, trajectories diverged in a way consistent with a bias toward the reference. This effect spanned 54 successive time steps with significant differences at $p < .01$, as well exceeding the bootstrap criterion of 6 time steps. For this effect, the sequence of significant differences extended from 64.9 to 100 percent of movement time, with the minimum p value occurring at 100 percent movement time ($t(11) = 4.43$, $p < .01$, $d_z = 1.28$).

For horizontal axis spatial terms (Figure 3.11e), only the early divergence consistent with a bias away from the reference remained. Note that, due to the coupling of reference side and spatial term in trials with horizontal axis spatial terms, this bias is also congruent with movement in the direction described by the spatial term. The divergence was present over 64 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 34 time steps. The sequence of significant differences extended from 1.32 to 43.05 percent of movement time, with the minimum p value occurring at 1.99 percent movement time ($t(11) = -11.15$, $p < .001$, $d_z = -3.22$).

For vertical axis spatial terms (Figure 3.11f), in contrast, only the late divergence consistent with a bias toward the reference remained. The divergence was present over 103 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 21 time steps. The sequence of significant differences extended from 32.45 to 100 percent of movement time, with the minimum p value occurring at 78.15 percent movement time ($t(11) = 12.82$, $p < .001$, $d_z = 3.7$).

3.1.3 Brief discussion

In the vast majority of trials, participants selected the target item, that is, the item that had the target color specified in the spatial phrase and that provided the best match for the spatial term according to the spatial templates used for display creation. Thus, the employed spatial template shapes were appropriate for this purpose.

The mouse paths to the target item displayed biases into different direc-

tions. Three effects were observed, in line with the hypotheses.

First, there was a *distractor effect*, which biased trajectories to the side of the direct path on which the distractor item was located. The effect was observed to comparable degrees when the target position relative to the reference item was specified by horizontal axis spatial terms ("left of" or "right of") and when it was specified by vertical axis spatial terms ("above" or "below"). Its onset occurred after approximately a third of the total movement time.

Second, there was a *reference effect*, which consisted of trajectory attraction toward the side of the direct path where the reference item of the spatial phrase was located. In the mean data across spatial term axes this effect was visible within the last third of movement time. In trials using the spatial terms "above" and "below", its onset occurred after approximately a third of the total movement time, and the effect was considerably more pronounced, likely due to not being superimposed with the spatial term effect, discussed below. An attraction to the reference item was not observed for horizontal axis spatial terms, which again was probably due to superimposition with the spatial term effect in these trials. That the reference effect was weaker in the across-spatial term comparison is most likely attributable to the mixture of trials from both spatial term axes in that comparison, so that an average of the effect's presence in vertical axis spatial terms and its absence in horizontal axis spatial terms was observed.

Third, the *spatial term effect* consisted of a bias with very early onset, into the direction described by the spatial term. It was visible in the comparisons by reference side and there only in the mean data across spatial terms and, more strongly and somewhat more extended, in trials with horizontal axis spatial terms. In both cases, the effect occurred immediately after movement onset and remained observable over approximately 40 percent of the total movement time. Note that the spatial term used in a trial did not predict the location of the target or its side in the display, because each possible target position was used an equal number of times per spatial term.

That the spatial term effect was observed only in the comparisons between reference sides, and only for horizontal axis spatial terms (and less strongly in the across-spatial term comparison) is unsurprising. In trials with horizontal axis spatial terms, the side of the reference item relative to the direct path was coupled to the spatial term (see *Balancing the effects of potentially confounding items*, p.81): When the reference item was on the left side, the spatial term was "right of" and when the reference item was on the right side, the spatial term was "left of". Thus, in the comparisons by reference side for horizontal

axis spatial terms each set of data included only one spatial term, so that its effect could be seen in the mean trajectories.

In trials with vertical axis spatial terms, in contrast, reference sides were not coupled to specific spatial terms. Furthermore, any biases into the direction of vertical spatial terms would likely not have become observable in the trajectories, as such biases would have acted approximately parallel to the principal direction of movement and thus orthogonally to the axis along which deviation was assessed.

Similar to the reference effect, the intermediate strength and earlier offset of the spatial term effect in the across-spatial term comparison by reference side is likely attributable to averaging over trials in which the effect was present and trials in which it was absent. It is furthermore likely that both the reference effect and the spatial term effect did affect trajectories in the second half of the comparison by reference side for horizontal axis spatial terms, and did not become visible in the mean data due to being superimposed and thus canceling each other out. This is assumed because there is no theoretical reason to expect the principal absence of reference attraction in these trials. Thus, although not seen in the mean data, the bias into spatial term direction likely affected trajectories for much longer than half of the the total movement time.

For an analogous reason, the spatial term effect can most likely be attributed to the linguistic spatial term rather than representing a repulsion from the reference item. In the latter case, the effect would have been expected to also be observable for vertical axis spatial terms. In addition, the effect's very early onset, which coincides with display onset, suggests that participants tended to move into a direction congruent with the spatial term already before display onset, so that the effect cannot have resulted from the arrangement of visual items.

An unbiased picture of the approximate onset time of biases can be gathered from those conditions where observed biases were probably not mixtures of multiples effects; this includes all comparisons by distractor side and the comparison by reference side for vertical axis spatial terms. Apart from the spatial term effect, biases' onsets occurred at approximately 30 to 40 percent of total movement time. Given an overall mean movement time of 1073 milliseconds, this corresponds to an absolute temporal separation between display onset and effect onset of approximately four hundred milliseconds or less (note that this must be considered a rough estimate since potential associations of movement time and trajectory deviation may have an influence;

for instance, if effects were more pronounced in trials with higher movement time, the estimate would have to be increased). Effect sizes were broadly comparable across effects, with the most pronounced effect being the reference effect in the case of vertical axis spatial terms.

The distribution of curvature included a marginal number of strongly curved trajectories, which formed a relatively uniform long tail of the distribution, while the majority of trajectories were smoothly curved. This suggests that the vast majority of trajectories were subject to graded attraction, whereas decisions about motor targets may have been abruptly revised in only very few trials. In line with this, no bimodality was indicated for the distribution, suggesting that trajectory shapes were not governed by fundamentally different processes from trial to trial.

Together, effects observed in this experiment support the notion that allocating attention to different visual items in the course of spatial language grounding is reflected by directional biases in mouse movement trajectories.

3.2 Experiment two: Generalization over response metrics

The first goal of this experiment was to probe whether the three effects observed in experiment one would generalize to different response movement metrics. This would strengthen the notion that the effects were indeed based on the tight coupling of neural substrates involved in the grounding of spatial language to substrates in which motor goals compete, rather than arising from a peculiar interaction between the specific response parameters and task demands in experiment one.

The second goal of experiment two was to further examine the notion that the spatial term effect, observed in early trajectory portions in experiment one, was indeed based on the spatial term rather than a repulsion from the reference item.

The paradigm was almost identical to experiment one, with the main difference that responses were made along a horizontal rather than vertical axis, from the start marker on the left side of the screen to targets on the right side of the screen.

It was hypothesized that all effects from experiment one would occur in experiment two as well, in an analogous manner but with partly reversed couplings of spatial term axes and effects. First, as in experiment one, the

distractor effect was hypothesized to occur equally for both spatial term axes. Second, the spatial term effect was hypothesized to be observable as a bias in spatial term direction for vertical axis spatial terms but not for horizontal axis spatial terms. Conversely, the reference effect was hypothesized to be observable for horizontal axis spatial terms, but not for vertical axis spatial terms. In other words, the observable signatures of the reference effect and the spatial term effect were expected to be switched between spatial term axes compared to experiment one. These expectations rested on the assumption that the two biases in spatial term and reference direction would cancel each other out in trajectory portions where both were present, as assumed with regard to the results of experiment one.

This hypothesized pattern of results would argue for the generality of both the reference effect and the spatial term effect over spatial term axes and response axes.

3.2.1 Methods

Participants

Twenty-four participants (15 female, nine male) with a mean age of 25 years (SD = 4.1 years) were recruited by notices around the local campus, signed informed consent (Appendix B), and received €10 for participation. All participants were right-handed, as determined by the Edinburgh Handedness Inventory (Oldfield, 1971). The participants were naïve to the experimental hypotheses, native German speakers, had self-reported normal or corrected-to-normal vision, and no color vision deficiencies (Appendix A).

Procedure

Participants were instructed to start movement from the start marker into rightward direction (instead of upward). Each participant completed 465 trials (except for two, who completed 469, so that all all available trials were used). The slight difference in trial numbers to experiment one was due to a minor adjustment in the creation of visual displays, described in *Generating visual displays (p.102)*.

Material

Visual displays How the start marker and the item arrays were arranged on the screen is illustrated in Figure 3.12. The arrangement differed from that

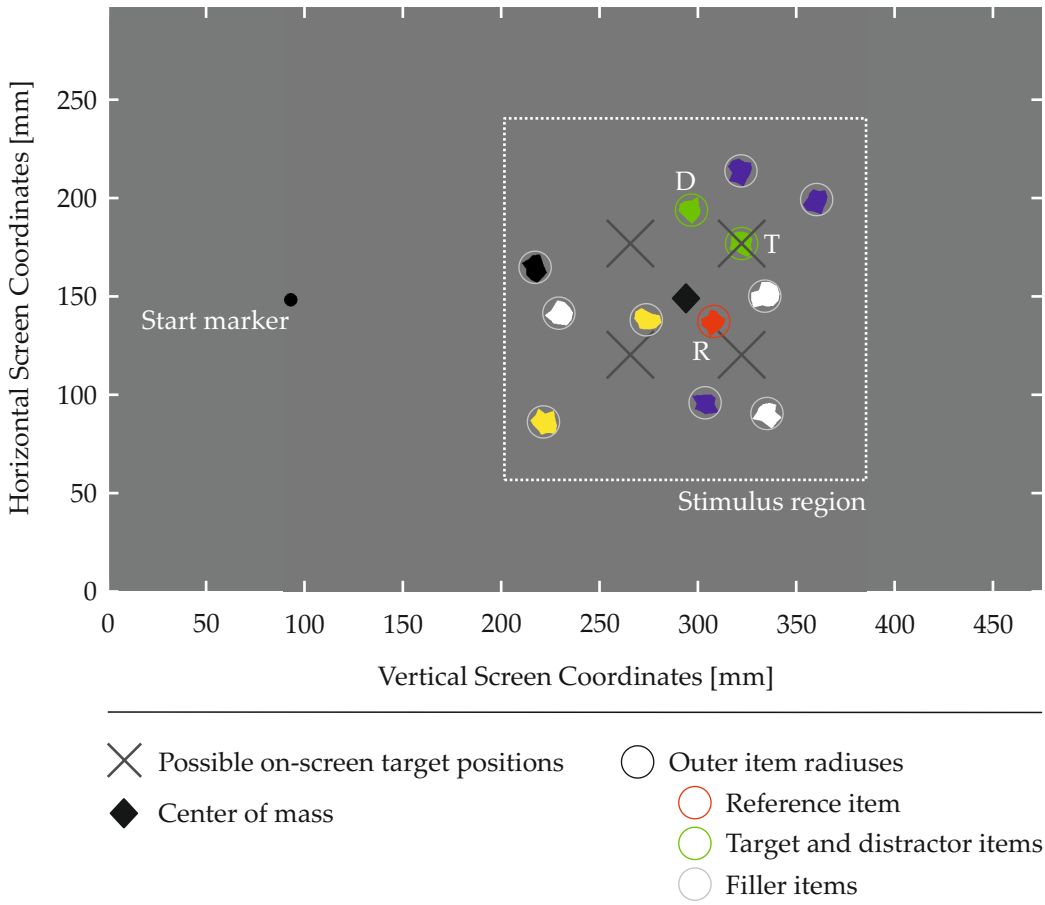


Figure 3.12: Display configuration in experiments two, three, and five. The shown area spanned the entire screen. The item configuration shown in the figure corresponds to the spatial phrase "The green one above the red one". T denotes the target, D the distractor, and R the reference item.

in experiment one only by clockwise rotation around the screen center by 90 degrees, whereas the extent of the different components and the distances between them were unchanged.

Generating visual displays Displays were created using the same methods as for experiment one. A minor change was that the placement of the grid of target positions over the target region was adjusted such that the possible target positions were distributed symmetrically on either side of the axis of the spatial term at hand (see Figure 3.13; this was the case for the remaining experiments as well). This change resulted in a slightly different number of distractor positions for each of 16 target positions, namely 19 to 24 per target with a mean of 21.8. In turn, this lead to a slightly higher total number of different displays, namely 5584. Since participant number was 24 and thus double the number of that in experiment one, the set of trials was doubled as well (i.e., each visual display was presented twice) and trials from the

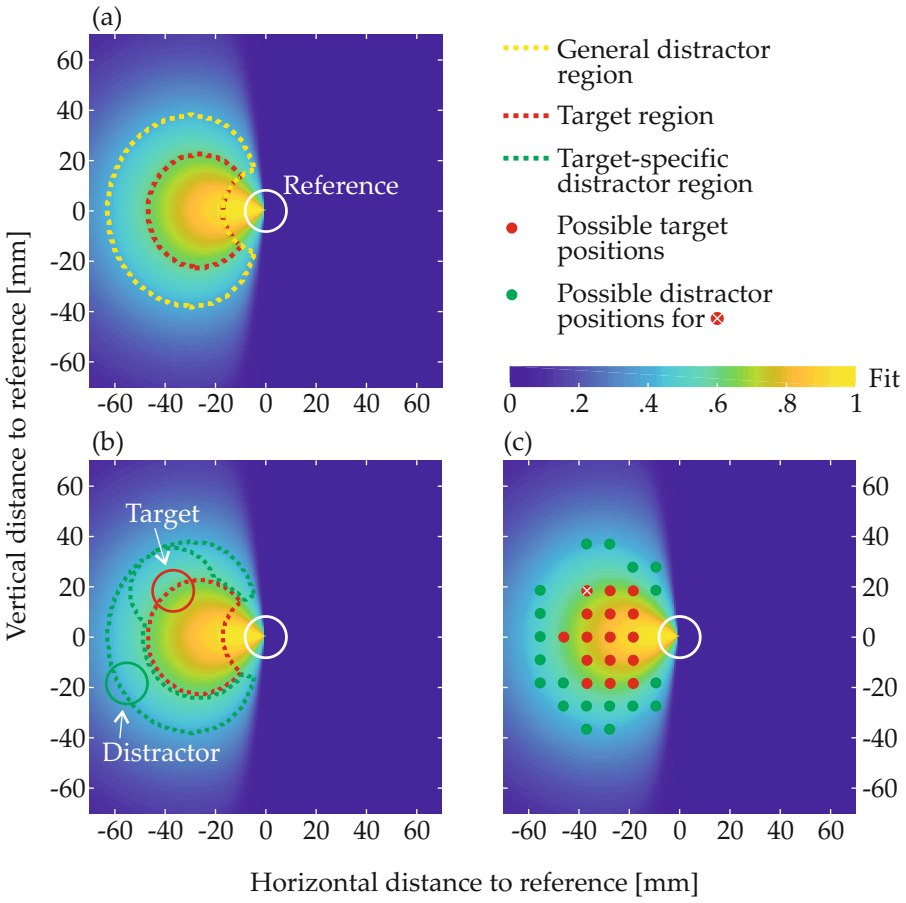


Figure 3.13: Item placement in experiment two (also used in experiments three to five) based on fit constraints, here using the spatial template for 'left of'. (a) Regions generally eligible for target and distractor placement defined only by fit cut-off and minimal item distance (item centers must be placed within these). (b) Exemplary placement of target and distractor item. Note how the distractor region for this specific target placement shrinks compared to (a). (c) All possible target positions for 'left of' and all possible distractor positions for the marked target position. Large circles depict outer radiuses of the items.

resulting set of size 11168 were randomly assigned to the participants.

Analysis

Analysis was the same as in experiment one. This largely extended to balancing methods, but note that the schemes underlying the composition of overall means for experiment two differed from those of experiment one (see *Balancing the effects of potentially confounding items*, p.81, Figure 3.8, 3.9). This was owed to the switched relationship between the principal movement direction and spatial term axes: in experiment two, the main axis of the terms 'left' and 'right' was approximately parallel to the direct paths and the axis of the spatial terms 'above' and 'below' was approximately orthogonal to the direct paths. Thus, constraints with respect to item placement and balancing categories that were associated with horizontal axis spatial terms in experiment

one were instead associated with vertical axis spatial terms in experiment two. This pertained to the side of the reference item relative to the direct path, which was fixed for each of the vertical axis spatial terms in experiment two (left side for 'right', right side for 'left'). In addition, the coupling of CoM side and on-screen target positions changed as well, in that the CoM was to the right of the direct path when one of the upper on-screen target positions was used and to the left of it when the lower ones were used.

The changed couplings in experiment two also affected the balancing categories that could be satisfied for each condition, with vertical axis spatial terms being more restricted in this respect in experiment two. The resulting balancing scheme for the distractor effect in experiment two is illustrated in Figure 3.14 and for the reference effect in Figure 3.15.

Figure 3.14 (on following page): Schematic depiction of possible item configurations in experiments two, three, and five for each combination of distractor side, on-screen target position, and spatial term. Note that displays are shown in vertical orientation due to space constraints but were presented rotated clockwise by 90 degrees. Balancing categories are indicated in the bottom right of each panel. Panels are arranged by distractor side to illustrate which conditions were combined and compared to examine the distractor effect. Means from panels sharing outline color and style were averaged, and the resulting means were compared between sets of differing outline style (within and across colors). Target (T) and distractor (D) are shown as green dots, the reference (R) is shown in red, and the center of mass across all items (fillers not shown) is depicted by a black diamond. The start marker is shown as a black dot and light gray crosses indicate possible on-screen target positions.

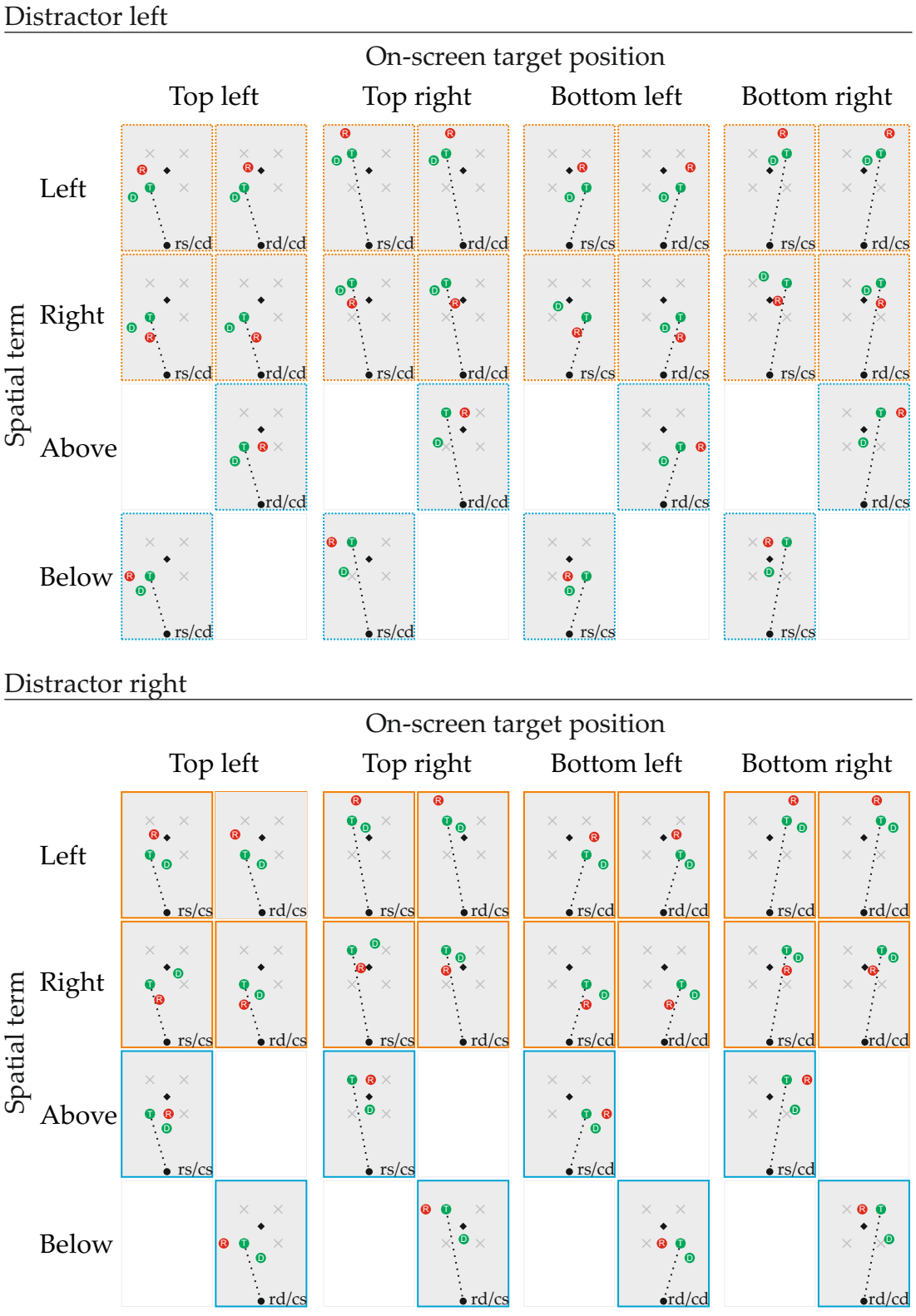


Figure 3.14: Caption on previous page.

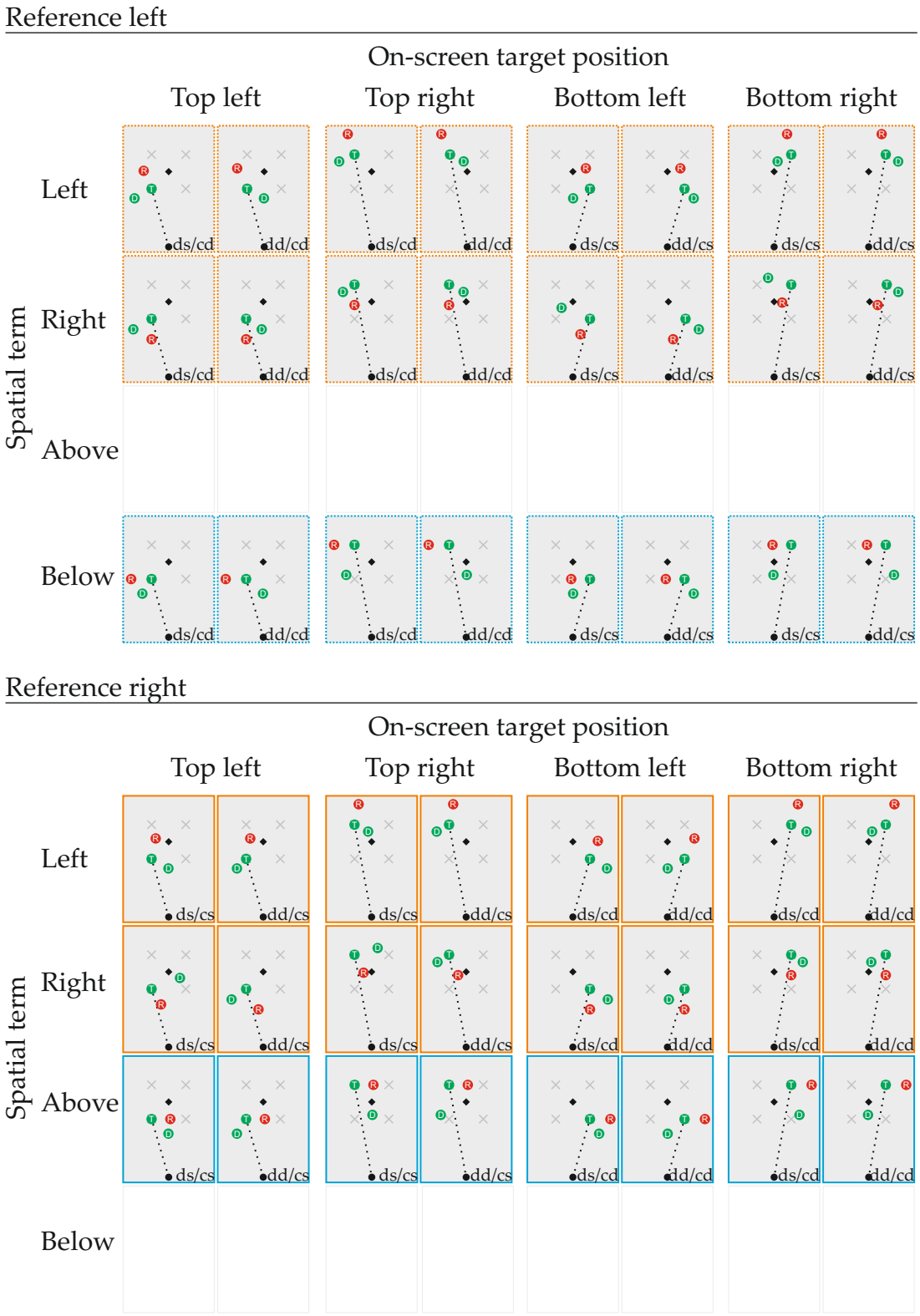


Figure 3.15: Caption on following page.

Figure 3.15 (on previous page): Schematic depiction of possible item configurations in experiments two, three, and five for each combination of reference side, on-screen target position, and spatial term. Note that displays are shown in vertical orientation due to space constraints but were presented rotated clockwise by 90 degrees. Balancing categories are indicated in the bottom right of each panel. Panels are arranged by reference side to illustrate which conditions were combined and compared to examine the reference effect. Means from panels sharing outline color and style were averaged, and the resulting means were compared between sets of differing outline style (within and across colors). Target (T) and distractor (D) are shown as green dots, the reference (R) is shown in red, and the center of mass across all items (fillers not shown) is depicted by a black diamond. The start marker is shown as a black dot and light gray crosses indicate possible on-screen target positions.

3.2.2 Results

When asked, participants reported not to have noticed that possible target positions were restricted to four on-screen locations. Movement onset was generally registered close to the center of the start marker (M = 2.15 mm, SD = 3.01 mm).

A total of 11157 trajectories was obtained (11 were lost due to technical issues). Of these, 10224 (91.64 %) were below curvature threshold (M = 426, SD = 32.9 equaling M = 91.64 %, SD = 7.09 %). Of the non-curved trajectories, 86.75 percent (8869) were correct responses and thus entered further analysis (79.49 % of all obtained trajectories).

Participants achieved a mean accuracy of 86.59 percent (SD = 5.68 %) and their mean movement time was 1057 milliseconds (SD = 129 ms). Note that the numbers reported so far were not affected by balancing but based on simple averaging over the respective trial ensembles; mean data reported from here on was obtained according to the balancing scheme described above.

Condition-specific movement times are listed in Table 3.5 and showed only marginal differences.

Figure 3.16 shows fifty examples for trajectories below and above curvature threshold (panels a and b, respectively) along with the empirical distribution over maximum curvature value for all correct responses (panel c),

Table 3.5: Movement times and standard deviations (SD) for experiment two.

	Distractor side				Reference side			
	Left		Right		Left		Right	
Spatial terms	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Overall	1061	127	1058	129	1058	123	1060	133
Left/Right	1066	130	1065	137	1064	132	1066	137
Above/Below	1058	121	1050	124	1055	114	1053	130

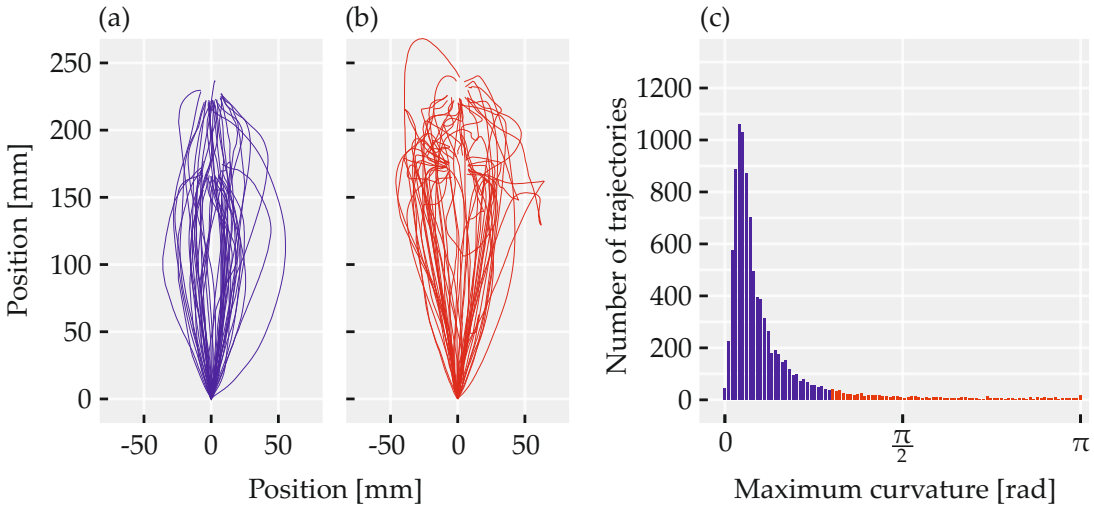


Figure 3.16: Fifty randomly selected trajectories obtained in experiment two that were (a) below curvature threshold and (b) exceeding curvature threshold. Panel c shows the overall distribution of trajectories over maximum curvature values, where red bars correspond to trajectories that were discarded due to high curvature. Only correct responses are shown.

with red bars indicating curvature above threshold (i.e., trajectories excluded from other analyses). For the distribution Hartigan's dip test indicated no bimodality ($p > .05$).

The top row of panels in Figure 3.17 visualizes the results of comparisons of mean trajectories by distractor side, where red and blue circles labeled 'D' in the top of each panel indicate distractor side for the correspondingly colored mean trajectory.

Across spatial terms, trajectories diverged in a way consistent with a bias toward the distractor (Figure 3.17a), with 93 successive time steps showing significant differences at $p < .01$, thus exceeding the bootstrap criterion ($p < .01$) of 7 time steps. The sequence of significant differences extended from 39.07 to 100 percent of movement time, with the minimum p value occurring at 73.51 percent movement time ($t(23) = 11.9$, $p < .001$, $d_z = 2.43$).

For horizontal axis spatial terms, the bias toward the distractor was present as well (Figure 3.17b), with 83 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 6 time steps. The sequence of significant differences extended from 45.70 to 100 percent of movement time, with the minimum p value occurring at 82.12 percent movement time ($t(23) = 10.58$, $p < .001$, $d_z = 2.16$).

Similarly, for vertical axis spatial terms, the bias toward the distractor was present (Figure 3.17c), with 71 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 5 time steps. The sequence of significant differences extended from 53.64 to 100 percent of

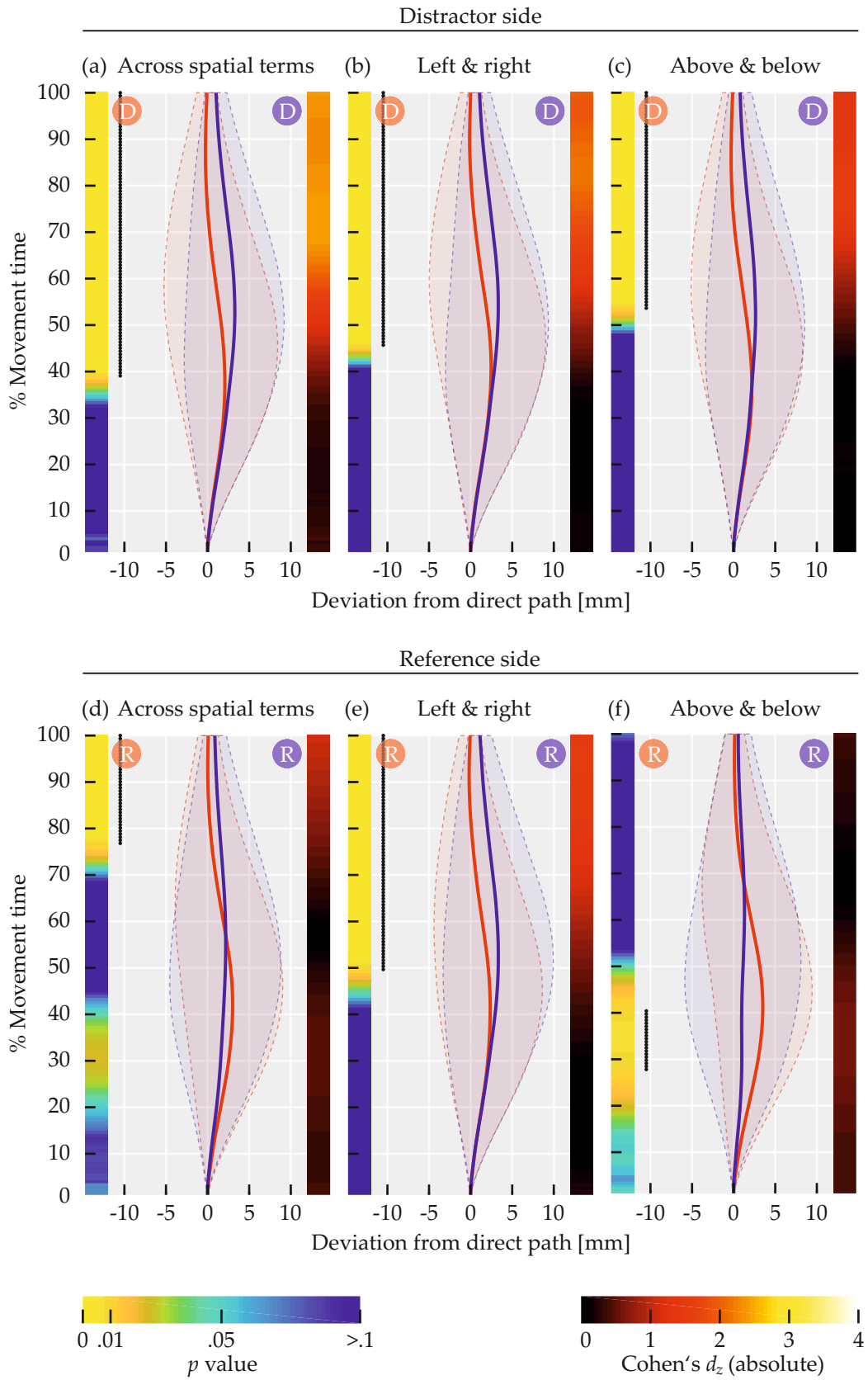


Figure 3.17: Comparisons of mean trajectories for experiment two. Red and blue circles labeled 'D' or 'R' indicate distractor or reference side, respectively, for the correspondingly colored mean trajectory. Transparent regions delimited by dashed lines indicate between-subjects standard deviation. Left image maps in the panels indicate p values from the t-tests at the time steps, right image maps indicate effect size (absolute Cohen's d_z). Black dotted lines on the left span time steps with significant differences between mean trajectories.

movement time, with the minimum p value occurring at 79.47 percent movement time ($t(23) = 7.59$, $p < .001$, $d_z = 1.55$).

The bottom row of panels in Figure 3.17 visualizes the results of the comparisons by reference side, where red and blue circles labeled 'R' in the top of each panel indicate reference side for the correspondingly colored mean trajectory.

Across spatial terms (Figure 3.17d), trajectories diverged in a way consistent with a bias toward the reference. This effect spanned 36 successive time steps with significant differences at $p < .01$, exceeding the bootstrap criterion of 4 time steps. For this effect, the sequence of significant differences extended from 76.82 to 100 percent of movement time, with the minimum p value occurring at 100 percent movement time ($t(23) = 5.43$, $p < .001$, $d_z = 1.11$). The expected bias in spatial term direction in the first half of movement time was visible only as a non-significant tendency (the minimum p value within the first half was $p = .023$ at 30.46 percent movement time; $t(23) = -2.44$).

For horizontal axis spatial terms (Figure 3.17e), the late divergence toward the reference item was more pronounced and became visible earlier earlier. It was present over 77 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 8 time steps. The sequence of significant differences extended from 49.67 to 100 percent of movement time, with the minimum p value occurring at 94.7 percent movement time ($t(23) = 7.46$, $p < .001$, $d_z = 1.52$).

For vertical axis spatial terms (Figure 3.17f), the early divergence consistent with a bias away from the reference (i.e., in spatial term direction) was significant. It was present over 20 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 3 time steps. The sequence of significant differences extended from 27.81 to 40.4 percent of movement time, with the minimum p value occurring at 36.42 percent movement time ($t(23) = -2.94$, $p < .01$, $d_z = -0.6$).

3.2.3 Brief discussion

The results of experiment two were largely analogous to those of experiment one. Mean accuracy was marginally lower, but the target item was still chosen in most trials. There were slightly more trajectories exceeding curvature threshold, potentially reducing the quality of estimated means due to less trials entering analyses. The distribution of curvature was not bimodal, however, and had a similar shape as in experiment one.

As hypothesized, a distractor effect was found in all comparisons by distractor side. The effect was overall comparable to its counterpart in experiment one, although its onset occurred somewhat later than before, and it was somewhat weaker in effect sizes and mean differences.

Furthermore, both a reference effect and a spatial term effect were found in comparisons by reference side. Most importantly, the hypothesized switch of the two effects with respect to the spatial term axes for which they occurred was observed. Attraction toward the side of the reference item was observed for horizontal axes spatial terms, as opposed to experiment one where this effect was seen for vertical and not for horizontal axes spatial terms. Conversely, an early bias into the direction described by the spatial term occurred for vertical axes spatial terms, as opposed to experiment one where this effect was seen for horizontal but not for vertical axes spatial terms.

This switch confirms, first, that the attraction to the reference item is a general effect not dependent on spatial term axes or response direction, and, second, that a spatial term effect is exerted by both types of spatial terms, those with a horizontal axis, and those with a vertical axis. The switch also confirms that the spatial term was not a repulsion from the reference item.

Note that both the onset time and the magnitude of the reference effect were reduced compared to experiment one; in experiment two, it became significant only after approximately half of the total movement time. A more pronounced difference to experiment one arose for the spatial term effect, which in experiment two became significant only shortly after movement onset (while its end occurred at a time more similar to experiment one). This also affected the comparison by reference side across spatial terms, where the spatial term effect did not become significant. However, a trend toward significance in the earlier portion of the comparison by reference side for vertical axis spatial terms was present, suggesting that there was no fundamental difference between the spatial term effect observed here and in experiment one.

Finally, the between-subjects standard deviation of trajectories was overall considerably larger than in experiment one, as obvious from comparing shaded regions in Figure 3.11 and Figure 3.17. The reason for this is unclear; the low variability in experiment one may have been a pattern based on chance, especially given the relatively low number of participants in experiment one (and because subsequent experiments showed a degree of between-subjects variability of mean trajectories that was more similar to experiment two than experiment one). Note that what mattered for the statistical paired-sample tests was not the between-subjects variability of compared mean tra-

jectories, but the between-subjects variability of difference scores between compared mean trajectories. This latter variability did not differ markedly in experiment one and two (as can be derived from the figures by relating the difference between mean trajectories to effect size).

In summary, although the effects were somewhat weaker in experiment two, they were generally in line with experiment one. The hypotheses were borne out, suggesting that the three effects generalize to different response metrics. This lends further support to the notion that the reference effect and the distractor effect were based on attraction toward visual items involved in spatial language grounding, and that the spatial term effect represented an influence of the semantics of linguistic spatial terms on motor action.

3.3 Experiment three: The effect of mouse movement speed

Experiment three probed whether the effects observed so far generalized to a higher gain mapping between mouse movement and cursor movement, as used in other mouse tracking research.

The mapping used in experiment one and two (and in the following ones), where movement on the desk translated to the same movement distance on the screen, was chosen to make the required arm movements more similar to natural reaching, and to simplify the necessary coordinate transforms between the two spaces. This was owed to the general focus on showing the embodiment of relation grounding and the modal nature of the underlying representations through their link to the motor level. The underlying rationale for using a low mouse gain was that behavioral signatures would be more likely to be discernible with fewer or less complex intervening processes.

However, this gives rise to the question how the trajectories obtained with this method relate to previous research and whether different mouse gains are able to show similar effects. Unfortunately, most mouse tracking studies do not report the employed desktop-to-screen mapping. The commonly used Mouse Tracker software (Freeman & Ambady, 2010) allows to influence the mapping only through an optional parameter that must be added manually and which is set in units of mouse speed under Microsoft Windows. How exactly these translate to the ratio of movement on the desk to movement of the cursor on the screen is not fixed, however, but depends on technical specifications of the hardware used (e.g., the sensitivity of the mouse's sensor).

It was therefore assumed that previous studies have used a mapping that is typical for mouses on desktop computers, and which is characterized by a higher gain than that used in the remaining experiments here.

The tentative and conservative hypothesis was that effects would be qualitatively similar to those observed in experiment two. Shorter movement times due to the faster mapping and a resulting shift of effect signatures in normalized time were expected, but not cast into formal hypotheses. In a similarly exploratory manner, experiment three also served as a comparison of methods, a probe whether one type of mapping would be superior to the other in making effects of cognitive processing discernible in mean trajectories.

The experiment used the same trials as experiment two, meaning that the principal axis of response movements was horizontal. The only difference to experiment two was the higher gain of cursor movement.

3.3.1 Methods

Participants

Twelve participants (six female, six male) with a mean age of 24.2 years (SD = 4 years) were recruited by notices around the local campus, signed informed consent (Appendix B), and received €10 for participation. All participants were right-handed, as determined by the Edinburgh Handedness Inventory (Oldfield, 1971). The participants were naïve to the experimental hypotheses, native German speakers, had self-reported normal or corrected-to-normal vision, and no color vision deficiencies (Appendix A).

Material

In contrast to all other experiments, mouse speed was set such that movement on the tabletop translated to cursor movement on the screen by a one-to-five ratio (one centimeter movement on the desktop translated to five centimeters on the screen). Trials were identical to experiment two, but each trial was presented only once in experiment three, resulting in 5584 trials in total.

3.3.2 Results

A total of 5584 trajectories was obtained, 3615 (64.74 %) of which were below curvature threshold (M = 301.25, SD = 82.91 equaling M = 64.73 %, SD = 17.79 %). Of the non-curved trajectories, 80.8 percent (2921) were correct

Table 3.6: Movement times and standard deviations (SD) for experiment three.

	Distractor side				Reference side			
	Left		Right		Left		Right	
Spatial terms	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Overall	758	100	768	93	762	105	764	90
Left/Right	775	129	775	112	766	126	781	107
Above/Below	756	98	764	90	766	98	754	89

responses and thus entered further analysis (52.31 % of all obtained trajectories). Movement onset was generally registered close to the center of the start marker (M = 1.82 mm, SD = 2.96 mm).

Participants achieved a mean accuracy of 81 percent (SD = 10.06 %) and their mean movement time was 760 milliseconds (SD = 92 ms). Note that the numbers reported so far were not affected by balancing but based on simple averaging over the respective trial ensembles; mean data reported from here on was obtained according to the balancing scheme described above.

Condition-specific movement times are listed in Table 3.6 and showed only marginal differences but were considerably lower than for the preceding experiments.

Figure 3.18 shows fifty examples for trajectories below and above curvature threshold (panels a and b, respectively) along with the empirical distribution over maximum curvature value for all correct responses (panel c), with red bars indicating curvature above threshold (i.e., trajectories excluded from other analyses). For the distribution Hartigan's dip test indicated bimodality ($p < .05$). With trajectories exceeding curvature threshold excluded, Hartigan's dip test indicated no bimodality ($p > .05$).

The top row of panels in Figure 3.19 visualizes the results of comparisons of mean trajectories by distractor side, where red and blue circles labeled 'D' in the top of each panel indicate distractor side for the correspondingly colored mean trajectory.

Across spatial terms, trajectories diverged in a way consistent with a bias toward the distractor (Figure 3.19a), with 59 successive time steps showing significant differences at $p < .01$, thus exceeding the bootstrap criterion ($p < .01$) of 4 time steps. The sequence of significant differences extended from 61.59 to 100 percent of movement time, with the minimum p value occurring at 98.68 percent movement time ($t(11) = 11.8$, $p < .001$, $d_z = 3.41$).

For horizontal axis spatial terms (Figure 3.19b), there was only a minimal,

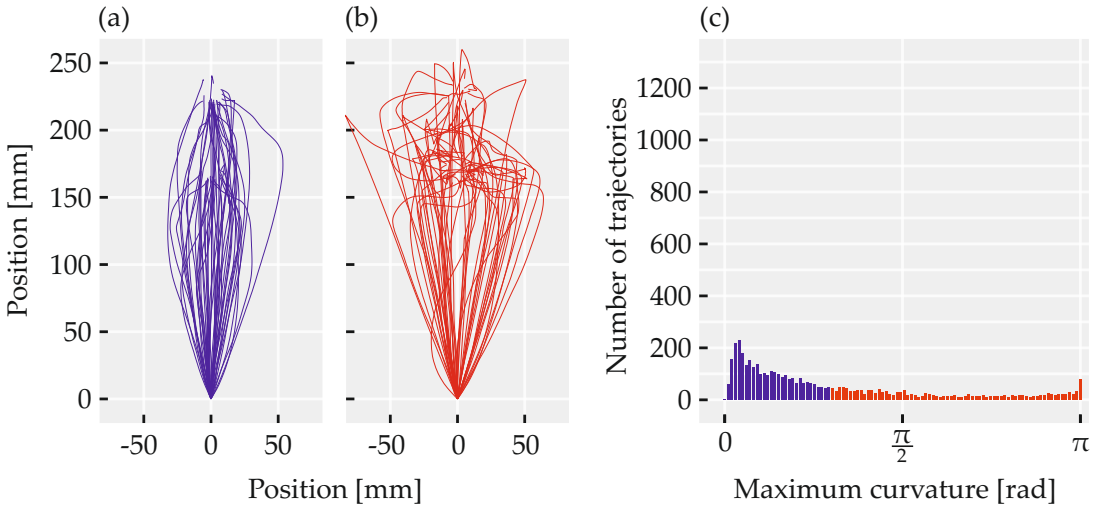


Figure 3.18: Fifty randomly selected trajectories obtained in experiment three that were (a) below curvature threshold and (b) exceeding curvature threshold. Panel c shows the overall distribution of trajectories over maximum curvature values, where red bars correspond to trajectories that were discarded due to high curvature. Only correct responses are shown.

non-significant tendency toward distractor attraction.

For vertical axis spatial terms (Figure 3.19c), in contrast, the bias toward the distractor was present, with 38 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 5 time steps. The sequence of significant differences extended from 75.50 to 100 percent of movement time, with the minimum p value occurring at 98.01 percent movement time ($t(11) = 5.79$, $p < .001$, $d_z = 1.67$).

The bottom row of panels in Figure 3.19 visualizes the results of the comparisons by reference side, where red and blue circles labeled 'R' in the top of each panel indicate reference side for the correspondingly colored mean trajectory.

Across spatial terms (Figure 3.19d), there were no significant differences between trajectories. However, there was a slight non-significant trend toward a bias in spatial term direction in the first half of movement time, with the minimum p value occurring at 34.44 percent movement time ($t(11) = -2.81$, $p = .0168$, $d_z = -0.81$). The expected bias toward the reference was non-significant and almost entirely absent, apart from a slight non-significant divergence at the end of movement time.

In contrast to this, for horizontal axis spatial terms (Figure 3.19e), trajectories diverged significantly in a way consistent with a bias toward the reference, although later than observed before. This effect spanned 21 successive time steps with significant differences at $p < .01$, exceeding the bootstrap criterion of 3 time steps. For this effect, the sequence of significant differences extended

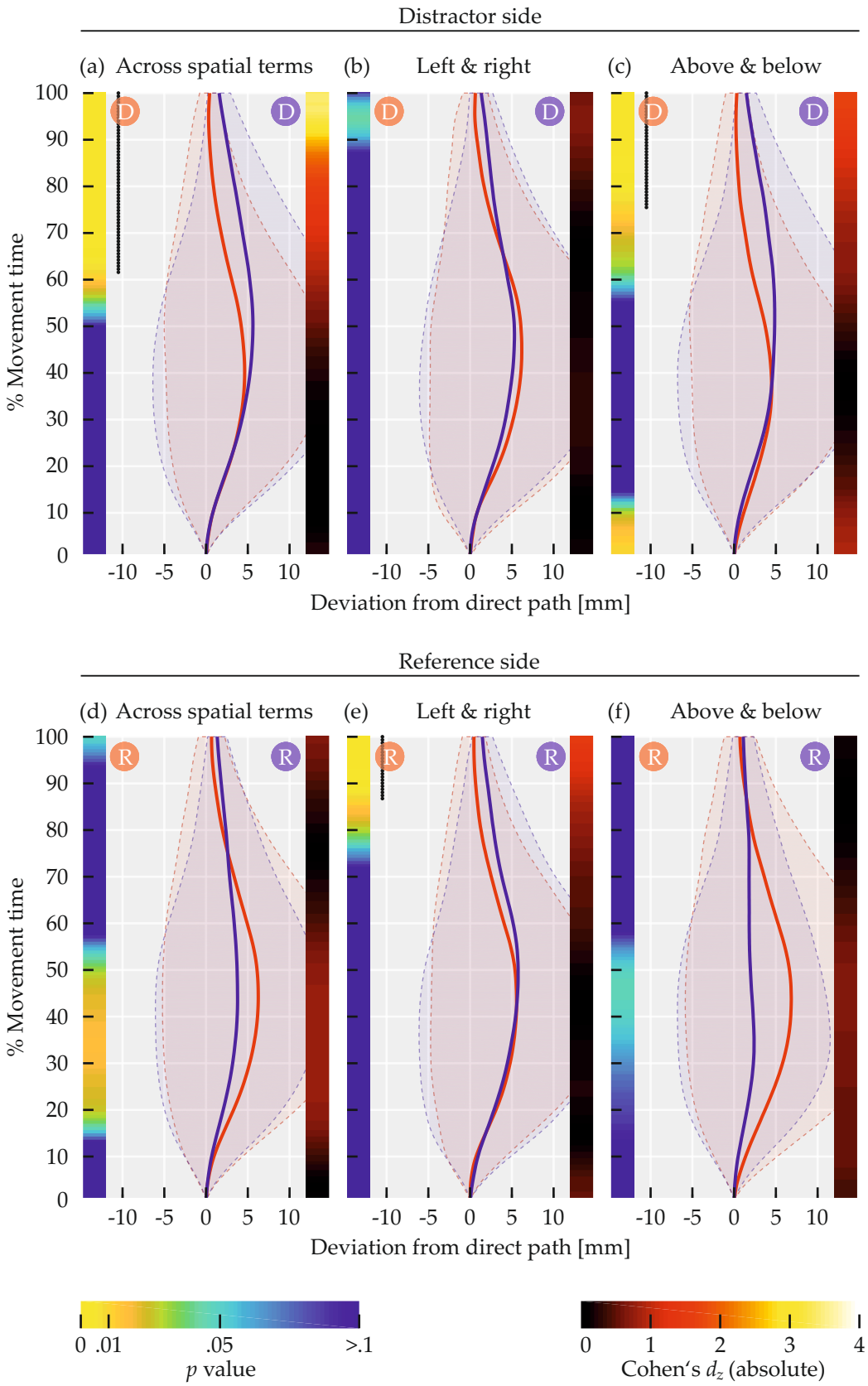


Figure 3.19: Comparisons of mean trajectories for experiment three. Red and blue circles labeled 'D' or 'R' indicate distractor or reference side for the correspondingly colored mean trajectory. Transparent regions delimited by dashed lines indicate between-subjects standard deviation. Image maps on the left side of each panel indicate the p value from the t-test at that time step, those on the right side indicate effect size (absolute Cohen's d_z). Black dotted lines on the left span time steps with significant differences between the mean trajectories.

from 86.75 to 100 percent of movement time, with the minimum p value occurring at 100 percent movement time ($t(11) = 5.34$, $p < .001$, $d_z = 1.54$).

For vertical axis spatial terms (Figure 3.19f), the expected divergence in spatial term direction was not significant, although hinted at by a weak trend (minimum p value occurring at 45.7 percent movement time with $t(11) = -2.26$, $p = .045$, $d_z = -0.65$).

3.3.3 Brief discussion

A high number of sharply curved trajectories lead to more than a third of all trajectories being excluded from analyses. This potentially reduced the quality of estimated means, which may have lead to effects being more difficult to discern. Overall variability in the remaining trajectories was larger than in previous experiments. As expected, movement times were considerably lower than in the previous experiments.

A distractor effect was found in the comparison across spatial terms and for vertical axis spatial terms, but not in horizontal axis spatial terms. Where the effect was present, the degree to which mean trajectories diverged and peak effect size were comparable to experiment two (peak effect size was somewhat stronger here). Why no effect was observed for horizontal axis spatial terms is unclear but may be connected to means being based on lower numbers of trajectories and to the generally higher variability.

The onset of the distractor effects occurred at a larger portion of total movement time than in previous experiments, that is, at about two thirds of movement time. Relating this onset in normalized time to an average movement time of approximately 760 milliseconds yields an (again, rough) estimate for absolute time between movement onset and effect onset of approximately 500 milliseconds. This is broadly consistent with what was derived for experiments one and two, albeit slightly longer.

No spatial term effect was observed. The reason for this is unclear, but given that trends hinting at such an effect were present in the respective comparisons, the reason may again be connected to high variability and the small number of trajectories entering analyses.

The reference effect was observed only for horizontal axis spatial terms and not in the overall comparison. Similar to the distractor effect, it became significant only toward the end of the total movement time. Its peak effect size and the size of the difference between mean trajectories was similar to that seen in experiment two.

As stated initially, many trajectories showed very high curvature, often close to that indicating antiparallel segments of trajectories (i.e., turns of 180°). Bimodality was found for the curvature distribution of all trajectories, likely based on a second peak located around the highest possible curvature value, but was removed in the set with those trajectories discarded that exceeded curvature threshold.

A possible reason for the large amount of curved trajectories may be linked to the overall short movement times. Where effects were found, they tended to occur around late portions of total movement time. This may hint that in many trajectories item positions could not affect the trajectories before the mouse cursor arrived in the vicinity of the item array. Participants may have moved toward the item array quickly, starting to ground the spatial phrase only when the cursor was in its vicinity. This would mean that the default movement direction must be corrected toward the target late in the overall trajectory, leading to a sharp turn.

Another hint at this explanation is that, for discarded trajectories (Figure 3.18), movement paths appeared to be relatively straight over large portions, while abrupt turns occurred only late, often resembling overshoots of mouse movement. According to this interpretation, the sharply curved trajectories emerged not due to successive cognitive processes in which first one, and then another item was selected for response in a discrete manner.

With regard to an explanation why the issue of overshoots seemed to affect the current study but not previous ones, there are two relevant properties of the current task that distinguish it from most other mouse tracking studies (see *Probing embodiment with mouse tracking*, p.16).

First, previous studies separated response options by wide angles, mostly placing the response buttons in the top corners of the screen and the starting position in its bottom center. This may discourage participants from moving quickly into a default direction, such as the screen center, as it decreases distance to the response options only marginally. In contrast, items were located in a relatively dense array in the paradigm used here, which affords quick movement into its general direction, as this will decrease distance to the target item independent of its exact location.

Second, in the paradigm used here, item presentation was triggered by movement onset, and the task instructions given to participants consequently facilitated initial movement into the general direction of the item array without information about the location of potential targets. Other mouse tracking studies typically showed the response options before movement onset, so that

it was not mandatory for participants to move into an intermediate direction first.

Importantly, the fact that expected effects were (partly) visible in the analyzed smoothly curved trajectories indicates that a late and abrupt correction of movement direction was not required for the effects to occur.

Together, these considerations suggest, first, that the high proportion of sharply curved trajectories seen in experiment three was at least partly due to mouse overshoots, and, second, that it is important to adjust the gain of cursor movement speed to other parameters of the experimental task at hand when departing from the canonical procedure. The alternative explanation for the bimodal distribution over curvature is that participants initially chose an item based on processes other than grounding the phrase, for instance, randomly or in some strategic manner, and then revised that decision in a similarly discrete manner, switching abruptly to the target as movement goal. Although the above considerations suggest otherwise, this explanation cannot be ruled out completely based on the present data.

Overall, the data obtained in experiment three likely suffered from the data loss due to discarded trajectories, but the distractor and the reference effect were found at least in some cases, as well as the expected shift in normalized time (this was interpreted as indicating that a relatively fixed amount of absolute time is required for effects of grounding to impact trajectories). Thus, the effects proved to generalize to this setup as well, although not as robustly as seen in the previous experiments. An insight emerging from comparison of these results to experiment one and two suggests that different mouse gains may be able to show similar effects, but that care must be taken to adjust gain to the other motor and cognitive parameters of the experimental task.

3.4 Experiment four: Exploring word order effects (vertical motion)

The goal of experiment four was twofold: It explored the impact of word order in the spatial phrases on the effects observed so far and it posed a replication of experiment one based on a higher number of participants.

To examine the impact of word order, half of all trials used the same spatial phrases as before while the other half used spatial phrases with modified word order (see below). Importantly, the data with respect to word order must be considered exploratory, because word order was added as a within-

subjects factor without increasing the total number of trials per participant (to avoid fatigue and compliance issues). For some conditions this resulted in few responses and thus low quality means, and where it lead to zero cases within one or more balancing categories jeopardized the balanced composition of overall means.

The approach to replicating the findings of experiment one depended on the presence or absence of a word order effect. If word order would prove to have no discernible effect, it was planned to collapse data across word order and analyze effects in the whole data set (negating the issues described above). If a word order effect would be present, replication would be restricted to the word order condition with the same phrases as those used so far.

The item arrays used in experiment four were the same as in previous experiments, but each was presented twice, once paired with the same spatial phrases as in previous experiments, and once paired with phrases of the form "Links vom Roten das Grüne" ("To the left of the red one the green one"). Thus, the new phrases described the same relations but the spatial term was mentioned first, followed by the reference item, and the target item in the last position. The word order from experiment one was labeled *TSR*, for target–spatial term–reference, and the new word order was labeled *SRT*, for spatial term–reference–target.

The inspiration to explore the impact of word order came from the visual world paradigm, which has shown that eye movements between items in a visual scene are closely time-locked to the corresponding words in concurrently perceived speech (Cooper, 1974; Tanenhaus et al., 1995; for review, see Salverda & Tanenhaus, 2017). The current paradigm is partly analogous to this, in that multiple visual items needed to receive attentional allocation, which in part was assumed to occur sequentially, in order to match each word to its referent in the scene and thereby derive an interpretation of the language input as a whole, while taking into account the constraints posed by the visual scene. An important difference is that in the current paradigm language was presented in written form, not verbally, and preceding the visual scenes, not concurrently.

Because linguistic input was fully known in advance, there were two possible sources of control for the sequence in which the different components were grounded: it could be governed either by computational properties of the underlying system (for instance, always grounding the reference item first), or by the ordered structure of the original language input (for instance, ground-

ing the three components according to their order in the phrase). While the former would predict the effects observed so far to be largely unaffected by the form of the phrase, the latter would predict word order to affect the temporal relationship between effects. In other words, an impact of word order on attraction effects would suggest that the neural mechanisms of grounding are sensitive to the original order in which linguistic information was presented, hinting at a non-stereotypical sequence of processes when grounding a spatial relation.

In the model described in Section 2 the 'blue print' for the sensorimotor processes of grounding is stored in the architecture of discrete nodes and the sequential organization of their activation is governed by the nodes' fixed connectivity. Thus, the model does not predict a dependency of effects on word order (when the entire phrase is known in advance; on-line language input is not within the model's scope). A dependency of the effect's temporal properties on word order would thus suggest a model extension to be necessary in terms of explicitly coding the order in which words in the linguistic input are delivered in the discrete neural representations that guide activation in the modal substrates.

The particular fixed order of selection in the model was not derived from evidence. Therefore, the expectation with regard to the effect of word order was based on the tentative assumption that the sequence of activation of the spatial phrase components would indeed be affected by word order, in analogy to findings of the visual world paradigm, and that this would result in a shift of effects in normalized time for word order SRT compared to TSR. It was assumed that the impact of the three components in the spatial phrase would shift in time analogous to their shift within the spatial phrase. Therefore, the distractor effect was expected to occur earlier for SRT than for TSR, the spatial term effect was expected to begin later in SRT than TSR, and the reference effect as well was expected to occur later in SRT.

3.4.1 Methods

Participants

Twenty-four participants (17 female, seven male) with a mean age of 24.6 years (SD = 4.4 years) were recruited by notices around the local campus, signed informed consent (Appendix B), and received €10 for participation. All participants were right-handed, as determined by the Edinburgh Handedness Inventory (Oldfield, 1971). The participants were naïve to the experimental

hypotheses, native German speakers, had self-reported normal or corrected-to-normal vision, and no color vision deficiencies (Appendix A).

Material

Spatial phrases Half of all trials included the same spatial phrases as in the preceding experiments, which had the word order target–spatial term–reference (TSR). The other half used phrases constructed according to Table 3.7, which were of the form exemplified by "Links vom Roten das Grüne.", translating to "To the left of [the] red [one] the green [one]". These phrases started with the spatial relation ("Links vom"), which was followed by a nominalized color word specifying the reference item of the trial ("Roten"), the article ("das"), and another nominalized color word specifying the target item of the trial ("Grüne."). Word order for these phrases thus was spatial term–reference–target (SRT). Note that the word order SRT is the only possible alternative word order to TSR in which the components of the spatial phrases can be arranged that is not agrammatical, although being non-standard usage.

Trials using the two word orders were presented in a randomly mixed fashion. Every two trials that used the same item array but different word orders were assigned to the same participant.

Table 3.7: Additional spatial phrases used in experiment four (and five). Source sets list the different candidates that filled the respective slots to form different spatial phrases.

General form	article	target item	spatial term	reference item
Example	**"Links vom**	**Roten**	**das**	**Grüne."**
Source sets	Links vom Rechts vom	Roten Grünen Gelben Blauen Schwarzen Weißen	das	Rote Grüne Gelbe Blaue Schwarze Weiße
English translation				
Example	**"Left of [the]**	**red [one]**	**the**	**green [one]."**
Source sets	Left of [the] Right of [the]	red [one] green [one] yellow [one] blue [one] black [one] white [one]	the	red [one] green [one] yellow [one] blue [one] black [one] white [one]

Visual displays

The start marker and item arrays were placed on the screen as in experiment one, so that responses were oriented vertically, travelling from the bottom of the screen in upward direction. Thus, the displays were largely similar to those in experiment one, with the exception that possible target positions were symmetrical across the axis of the spatial term at hand, as in experiment two and three. In consequence, the number of different visual displays was identical to these rather than experiment one (5584 different displays and 11168 trials in total due to each being presented once with each word order).

Statistical analysis

To test for effects of reference side and distractor side the same analyses as for the preceding experiments were applied, comparing left and right conditions across spatial terms, for vertical axis spatial terms, and for horizontal axis spatial terms.

As described initially, the new factor *word order* included the levels TSR and SRT. Its impact on the effects of reference and distractor side was assessed by comparing right−left difference scores (trajectory divergence) between word orders at each time step, separately for the reference and the distractor effect. These comparisons as well used series of paired sample t-tests and bootstrapping confirmation, both with $p < .01$. Additional alpha correction was not performed beyond that used in the previous experiments due to the exploratory nature of the additional tests for word order effects.

3.4.2 Results

A total of 11168 trajectories was obtained, 10223 (91.54 %) of which were below curvature threshold (M = 425.96, SD = 24.75 equaling M = 91.54 %, SD = 5.17 %). Of the non-curved trajectories, 82.71 percent (8455) were correct responses and thus entered further analysis (75.71 % of all obtained trajectories). Movement onset was generally registered close to the center of the start marker (M = 1.77 mm, SD = 2.98 mm).

Participants achieved a mean accuracy of 82.71 percent (SD = 7.86 %) and their mean movement time was 1013 milliseconds (SD = 130 ms). Note that the numbers reported so far were not affected by balancing but based on simple averaging over the respective trial ensembles; mean data reported from here on was obtained according to the balancing scheme described above.

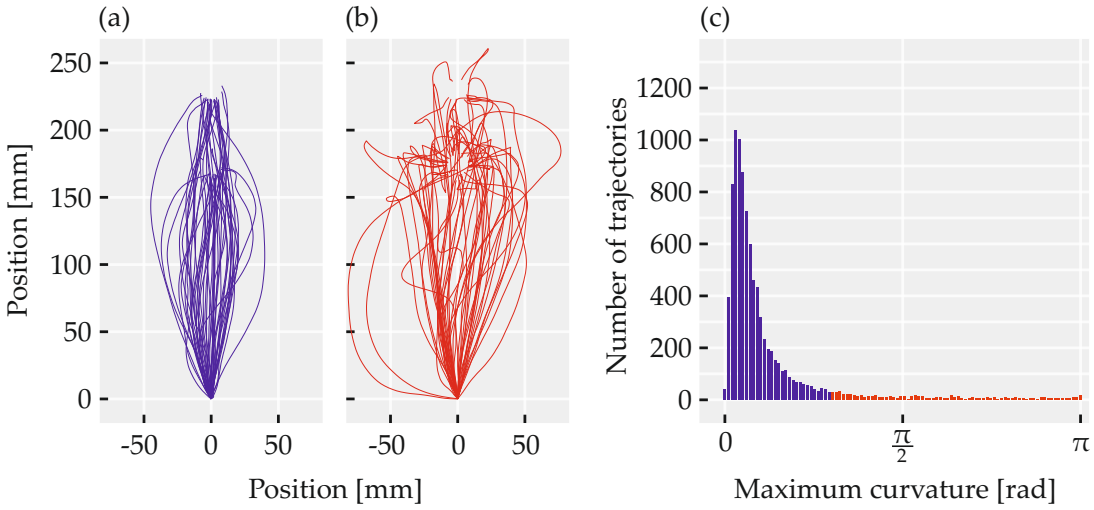


Figure 3.20: Fifty randomly selected trajectories obtained in experiment four that were (a) below curvature threshold and (b) exceeding curvature threshold. Panel c shows the overall distribution of trajectories over maximum curvature values, where red bars correspond to trajectories that were discarded due to high curvature. Only correct responses are shown.

Condition-specific movement times are listed in Table 3.8 and showed only marginal differences.

Figure 3.20 shows fifty examples for trajectories below and above curvature threshold (panels a and b, respectively) along with the empirical distribution over maximum curvature value for all correct responses (panel c), with red bars indicating curvature above threshold (i.e., trajectories excluded from other analyses). For the distribution Hartigan's dip test indicated no bimodality ($p > .05$).

The comparisons of trajectory divergence between word orders did not yield any significant effects, indicating that the impact of distractor and reference item side did not differ between TSR and SRT. The corresponding results will therefore not be discussed further; plots of the comparisons can be found in Appendix C. All data was collapsed across word order in order to examine it with respect to a replication of the effects seen in experiment one.

Table 3.8: Movement times and standard deviations (SD) for experiment four.

	Distractor side				Reference side			
	Left		Right		Left		Right	
Spatial terms	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Overall	1013	131	1019	131	1015	125	1018	137
Left/Right	1020	134	1031	141	1019	130	1032	144
Above/Below	993	130	1011	134	1007	128	997	135

The top row of panels in Figure 3.21 visualizes the results of comparisons of mean trajectories by distractor side, where red and blue circles labeled 'D' in the top of each panel indicate distractor side for the correspondingly colored mean trajectory.

Across spatial terms, trajectories diverged in a way consistent with a bias toward the distractor (Figure 3.21a), with 96 successive time steps showing significant differences at $p < .01$, thus exceeding the bootstrap criterion ($p < .01$) of 7 time steps. The sequence of significant differences extended from 37.09 to 100 percent of movement time, with the minimum p value occurring at 81.46 percent movement time ($t(23) = 7.59$, $p < .001$, $d_z = 1.55$).

For horizontal axis spatial terms, the bias toward the distractor was present as well (Figure 3.21b), with 74 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 5 time steps. The sequence of significant differences extended from 51.66 to 100 percent of movement time, with the minimum p value occurring at 80.13 percent movement time ($t(23) = 5.15$, $p < .001$, $d_z = 1.05$).

Similarly, for vertical axis spatial terms, the bias toward the distractor was present (Figure 3.21c), with 86 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 15 time steps. The sequence of significant differences extended from 43.71 to 100 percent of movement time, with the minimum p value occurring at 78.15 percent movement time ($t(23) = 6.85$, $p < .001$, $d_z = 1.4$).

The bottom row of panels in Figure 3.21 visualizes the results of the comparisons by reference side, where red and blue circles labeled 'R' in the top of each panel indicate reference side for the correspondingly colored mean trajectory.

Across spatial terms, a mixture of two biases was visible (Figure 3.21d). In the first half of movement time, trajectories diverged in a way consistent with a bias away from the reference item. This effect spanned 59 successive time steps with significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 16 time steps. For this effect, the sequence of significant differences extended from 1.32 to 39.74 percent of movement time, with the minimum p value occurring at 17.88 percent movement time ($t(23) = -6.63$, $p < .001$, $d_z = -1.35$).

In the second half, trajectories diverged in a way consistent with a bias toward the reference. This effect spanned 65 successive time steps with significant differences at $p < .01$, as well exceeding the bootstrap criterion of 16 time steps. For this effect, the sequence of significant differences extended

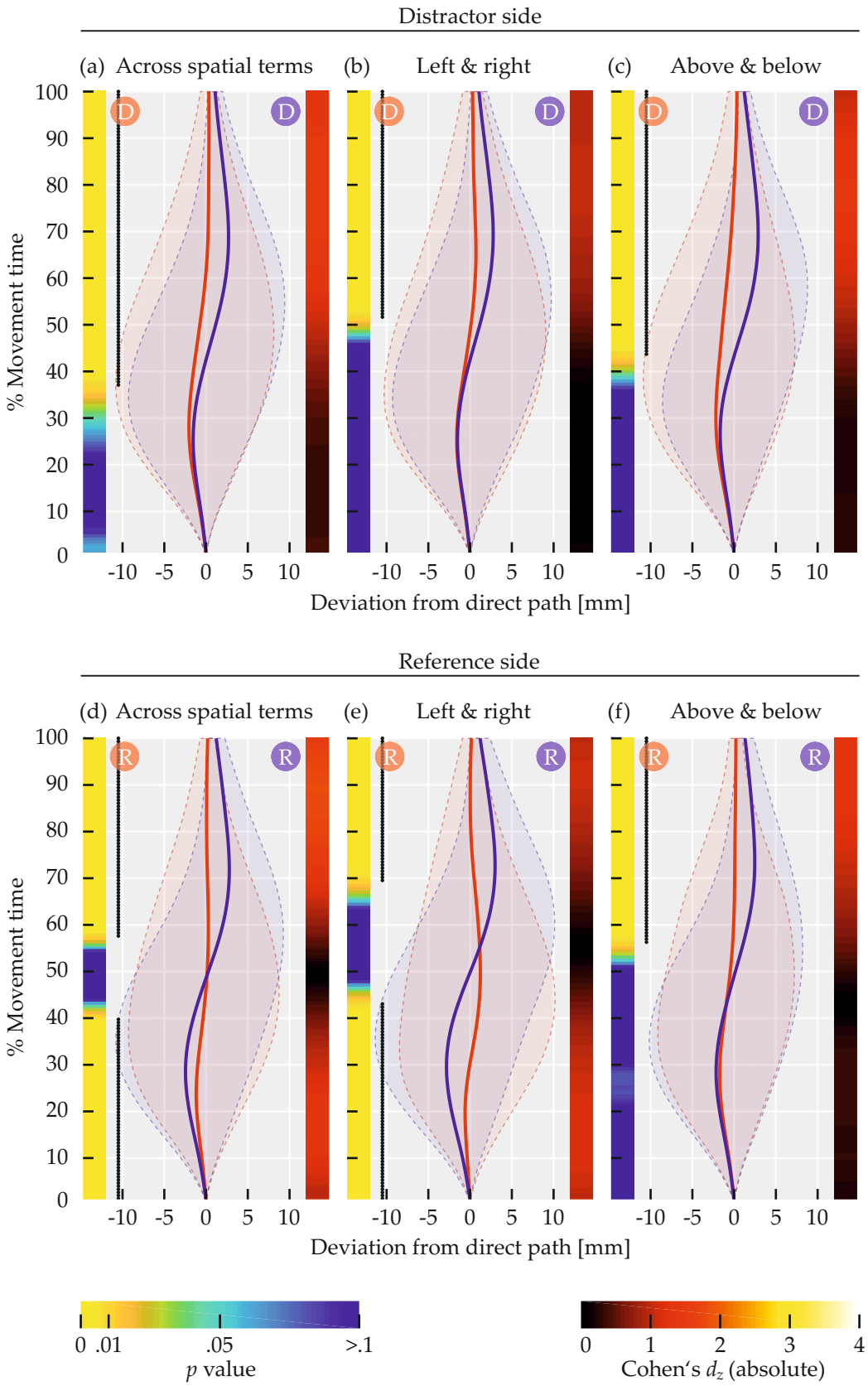


Figure 3.21: Comparisons of mean trajectories, across word order, for experiment four. Red and blue circles labeled 'D' or 'R' indicate distractor side or reference side, respectively, for the correspondingly colored mean trajectory. Transparent regions delimited by dashed lines indicate between-subjects standard deviation. Left image maps in panels indicate p values from t-tests at that time step, right image maps indicate effect sizes (absolute Cohen's d_z). Black dotted lines on the left span time steps where mean trajectories differ significantly.

from 57.62 to 100 percent of movement time, with the minimum p value occurring at 88.74 percent movement time ($t(23) = 8.89$, $p < .001$, $d_z = 1.82$).

The picture was very similar for horizontal axis spatial terms (Figure 3.21e), with both effects being visible. Here, the divergence in the first half, consistent with a bias in spatial term direction, was present over 64 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 12 time steps. The sequence of significant differences extended from 1.32 to 43.05 percent of movement time, with the minimum p value occurring at 17.88 percent movement time ($t(23) = -5.75$, $p < .001$, $d_z = -1.17$). The divergence in the second half, consistent with a bias toward the reference item, was present over 47 successive time steps showing significant differences at $p < .01$, as well exceeding the bootstrap criterion ($p < .01$) of 12 time steps. The sequence of significant differences extended from 69.54 to 100 percent of movement time, with the minimum p value occurring at 91.39 percent movement time ($t(23) = 5.3$, $p < .001$, $d_z = 1.08$).

For vertical axis spatial terms (Figure 3.21f), in contrast, only the late divergence consistent with a bias toward the reference remained. The divergence was present over 67 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 19 time steps. The sequence of significant differences extended from 56.92 to 100 percent of movement time, with the minimum p value occurring at 94.04 percent movement time ($t(23) = 6.53$, $p < .001$, $d_z = 1.33$).

3.4.3 Brief discussion

The expected effects of word order on trajectory divergence were not observed. However, as outlined initially, this data is of an exploratory nature. If consistent effects of word order had been found, they could have been interpreted more readily as showing that the order in which discrete conceptual information is conveyed in language guides grounding processes. Interpreting a null effect as showing the opposite is problematic, especially considering the exploratory nature of the data.

When considering the collapsed data, the findings of experiment one were largely replicated. A distractor effect was found in all comparisons by distractor side. Its magnitude with respect to both effect sizes and the difference between mean trajectories was reduced compared to experiment one, however, and was more similar to experiment two (note that variability as well was more similar to experiment two than experiment one, suggesting that the

low variability and particularly strong effects seen in experiment one were indeed a chance finding).

A similar picture emerged for the spatial term effect. It was very similar in onset and offset to experiment one, albeit somewhat weaker in effect size and with respect to the size of the difference between mean trajectories.

The reference effect was comparable to experiment one, but was weaker and showed a later onset (more similar to experiment two). However, in contrast to both experiment one and two, the reference effect was observed for both spatial term axes (and the across-spatial term comparison), again confirming that it represents a general effect of spatial terms.

Overall, experiment four provided inconclusive hints that word order in linguistic input does not guide the order of processes in spatial language grounding, and confirmed the findings of experiment one, while also being in accord with those of experiment two.

3.5 Experiment five: Exploring word order effects (horizontal motion)

Experiment five was identical to experiment four in all respects except that responses were performed by moving from left to right (as in experiment two and three). Thus, it posed a replication of experiment two in addition to exploring a possible impact of word order for the case of horizontally oriented responses.

3.5.1 Methods

Participants

Twenty-four participants (17 female, seven male) with a mean age of 24.4 years (SD = 4.1 years) were recruited by notices around the local campus, signed informed consent (Appendix B), and received €10 for participation. All participants were right-handed, as determined by the Edinburgh Handedness Inventory (Oldfield, 1971). The participants were naïve to the experimental hypotheses, native German speakers, had self-reported normal or corrected-to-normal vision, and no color vision deficiencies (Appendix A).

Table 3.9: Movement times and standard deviations (SD) for experiment five.

	Distractor side				Reference side			
	Left		Right		Left		Right	
Spatial terms	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Overall	1063	147	1055	151	1057	148	1062	150
Left/Right	1079	146	1061	154	1066	154	1074	149
Above/Below	1056	157	1054	150	1052	156	1058	151

3.5.2 Results

A total of 11168 trajectories was obtained. Of these, 9916 (88.79 %) were below curvature threshold (M = 413.17, SD = 40.5 equaling M = 88.79 %, SD = 8.66 %). Of the non-curved trajectories, 85.07 percent (8436) were correct responses and thus entered further analysis (75.54 % of all obtained trajectories). Movement onset was generally registered close to the center of the start marker (M = 1.93 mm, SD = 2.32 mm).

Participants achieved a mean accuracy of 84.94 percent (SD = 7.04 %) and their mean movement time was 1056 milliseconds (SD = 149 ms). Note that the numbers reported so far were not affected by balancing but based on simple averaging over the respective trial ensembles; mean data reported from here on was obtained according to the balancing scheme described above.

Condition-specific movement times are listed in Table 3.9 and showed only marginal differences.

Figure 3.22 shows fifty examples for trajectories below and above curvature threshold (panels a and b, respectively) along with the empirical distribution over maximum curvature value for all correct responses (panel c), with red bars indicating curvature above threshold (i.e., trajectories excluded from other analyses). For the distribution Hartigan's dip test indicated no bimodality ($p > .05$).

The comparisons of trajectory divergence between word orders did not yield any significant effects, indicating that the impact of distractor and reference item side did not differ between TSR and SRT. The corresponding results will therefore not be discussed further; plots of the comparisons can be found in Appendix D. All data was collapsed across word order in order to examine it with respect to a replication of the effects seen in experiment two.

The top row of panels in Figure 3.23 visualizes the results of comparisons of mean trajectories by distractor side, where red and blue circles labeled 'D' in the top of each panel indicate distractor side for the correspondingly

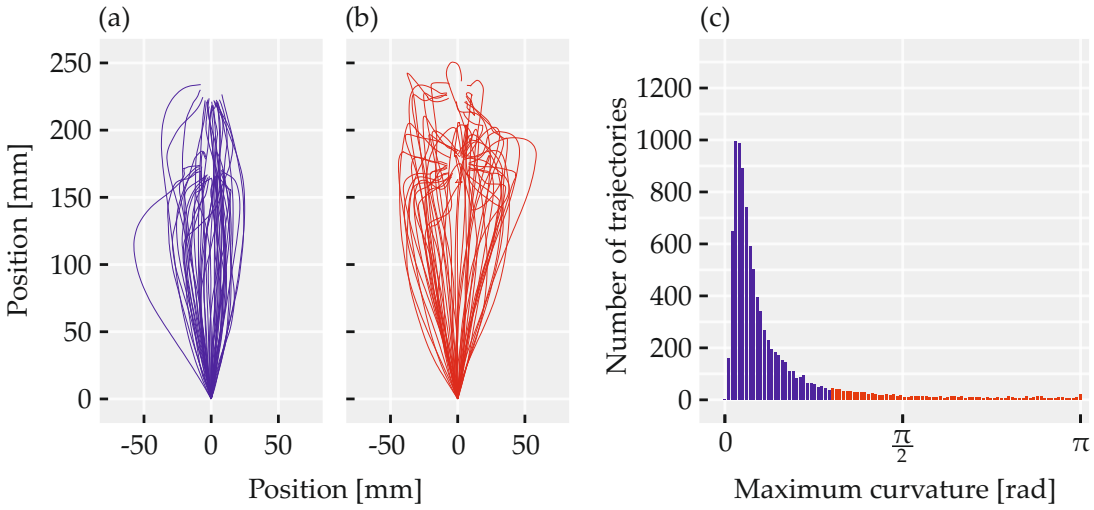


Figure 3.22: Fifty randomly selected trajectories obtained in experiment five that were (a) below curvature threshold and (b) exceeding curvature threshold. Panel c shows the overall distribution of trajectories over maximum curvature values, where red bars correspond to trajectories that were discarded due to high curvature. Only correct responses are shown.

colored mean trajectory.

Across spatial terms, trajectories diverged in a way consistent with a bias toward the distractor (Figure 3.23a), with 97 successive time steps showing significant differences at $p < .01$, thus exceeding the bootstrap criterion ($p < .01$) of 9 time steps. The sequence of significant differences extended from 36.42 to 100 percent of movement time, with the minimum p value occurring at 75.5 percent movement time ($t(23) = 10.94$, $p < .001$, $d_z = 2.23$).

For horizontal axis spatial terms, the bias toward the distractor was present as well (Figure 3.23b), with 56 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 4 time steps. The sequence of significant differences extended from 63.58 to 100 percent of movement time, with the minimum p value occurring at 86.09 percent movement time ($t(23) = 5.18$, $p < .001$, $d_z = 1.06$).

Similarly, for vertical axis spatial terms, the bias toward the distractor was present (Figure 3.23c), with 71 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 5 time steps. The sequence of significant differences extended from 53.64 to 100 percent of movement time, with the minimum p value occurring at 98.01 percent movement time ($t(23) = 6.54$, $p < .001$, $d_z = 1.33$).

The bottom row of panels in Figure 3.23 visualizes the results of the comparisons by reference side, where red and blue circles labeled 'R' in the top of each panel indicate reference side for the correspondingly colored mean trajectory.

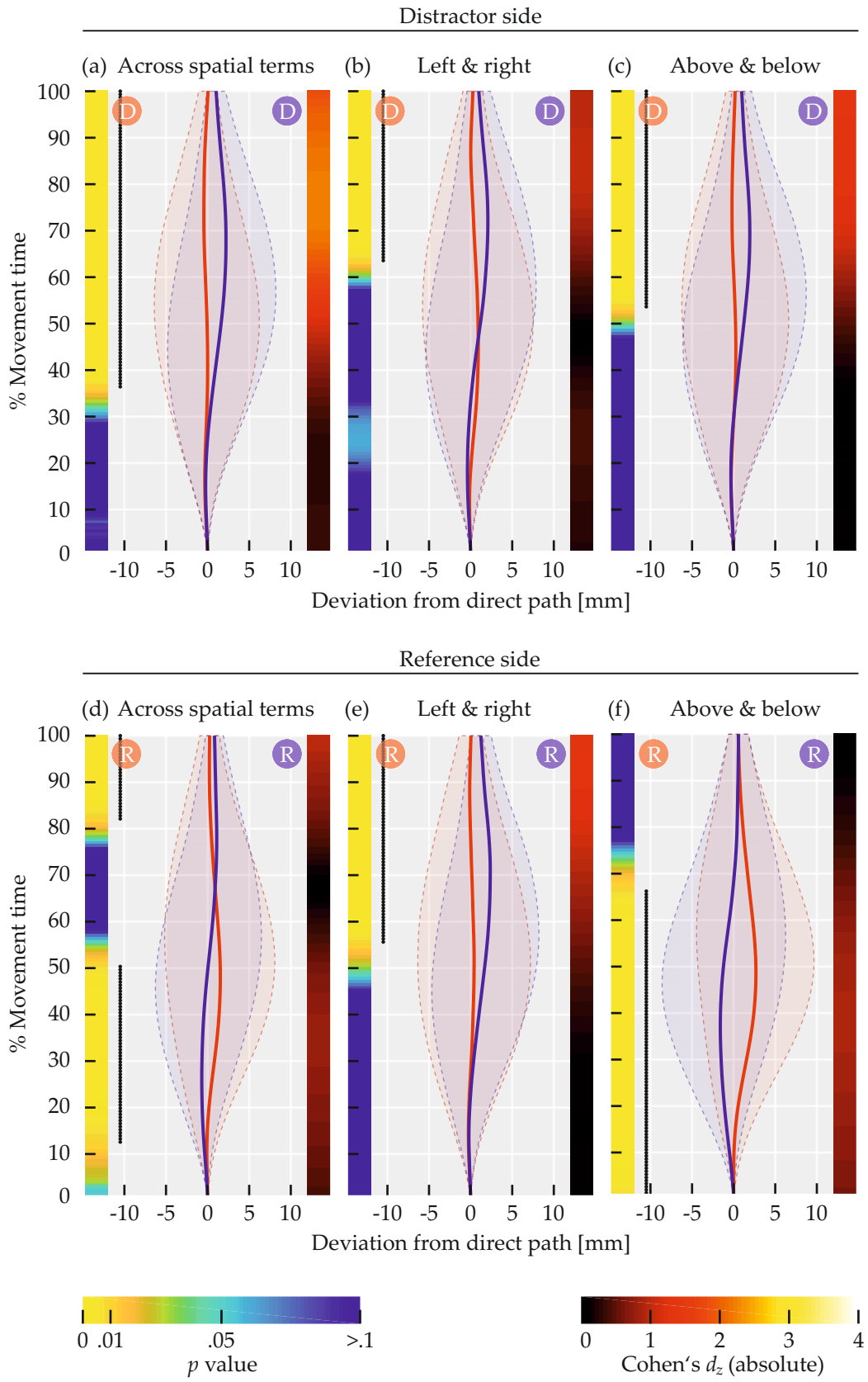


Figure 3.23: Comparisons of mean trajectory data, across word order, for experiment five. Red and blue circles labeled 'D' or 'R' in the top of each panel indicate distractor side or reference side, respectively, for the correspondingly colored mean trajectory. Transparent regions delimited by dashed lines indicate between-subjects standard deviation. Image maps on the left side of each panel indicate the p value from the t-test at that time step, those on the right side indicate effect size (absolute Cohen's d_z). Black dotted lines on the left span time steps with significant differences between the mean trajectories.

Across spatial terms, a mixture of two biases was visible (Figure 3.23d). In approximately the first half of movement time, trajectories diverged in a way consistent with a bias away from the reference item. This effect spanned 58 successive time steps with significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 3 time steps. For this effect, the sequence of significant differences extended from 12.58 to 50.33 percent of movement time, with the minimum p value occurring at 35.1 percent movement time ($t(23) = -3.8$, $p < .001$, $d_z = -0.78$).

In the second half, trajectories diverged in a way consistent with a bias toward the reference. This effect spanned 28 successive time steps with significant differences at $p < .01$, as well exceeding the bootstrap criterion of 3 time steps. For this effect, the sequence of significant differences extended from 82.12 to 100 percent of movement time, with the minimum p value occurring at 100 percent movement time ($t(23) = 4.92$, $p < .001$, $d_z = 1$).

For horizontal axis spatial terms (Figure 3.23e), only the late divergence toward the reference item remained and became visible earlier. It was present over 68 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 6 time steps. The sequence of significant differences extended from 55.63 to 100 percent of movement time, with the minimum p value occurring at 96.03 percent movement time ($t(23) = 6.42$, $p < .001$, $d_z = 1.31$).

For vertical axis spatial terms (Figure 3.23f), only the early divergence consistent with a bias away from the reference (i.e., in spatial term direction) was significant and further extended toward the end of the trajectory. It was present over 99 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 8 time steps. The sequence of significant differences extended from 1.32 to 66.23 percent of movement time, with the minimum p value occurring at 37.09 percent movement time ($t(23) = -4.38$, $p < .001$, $d_z = -0.9$).

3.5.3 Brief discussion

As in experiment four, no effect of word order on trajectory divergence was found.

The remaining results of experiment five replicated those of experiment two very closely, including the distractor effect and the reference effect. The only difference arose with respect to the spatial term effect, which for vertical axis spatial terms showed an early onset and extended over two thirds of

the total movement time and thus considerably further than in experiment two. This lends support to the notion that this bias may have affected large portions of trajectories in the previous experiments as well, rather than being restricted to early portions.

In summary, experiment five could not provide evidence for an impact of the order of words in language input on perceptual grounding processes, which is in line with experiment four. This can again not be taken as conclusive evidence on that matter due to the exploratory nature of the experiment in this respect. Other than that, findings of experiment two were largely replicated.

3.6 Experiment six: The effect of a competing relational pair

While the preceding experiments showed the biasing impact of individual visual items, this experiment focused on the effect of an additional relational pair. The goal was to investigate whether a biasing influence exerted by a second relational pair would transcend the sum of attraction caused by an additional reference item presented alone or by a distractor item presented alone. This would suggest that when items form a relational pair, a different or extended set of processes occurs than when items are presented in isolation.

In turn, this would support the general notion that attraction in this and the other experiments were signatures of flexible grounding processes that varied with the grounding scenario, as in the model described in Chapter 2. It would argue against stereotypical processes as the origin of the observed effects, such as focusing attention on each item a single time independent of the task scenario.

As before, visual displays included a relational pair, labeled *pair A* in the following, which instantiated the spatial term from the relational phrase and was composed of a reference item (*reference A*) and a target item. In addition, each display contained a second relational pair, called *pair B*, that was identical to pair A except for being flipped along the spatial term axis (Figure 3.24a). Pair B thus instantiated the reverse of the spatial term at hand. Thus, the item in pair B with the same color as the target item posed a *distractor*. The reference item in pair B will be referred to as *reference B*.

Two additional conditions were introduced that used the same visual displays as in trials with a full second pair, except that in one condition (*distractor*

only; Figure 3.24c) reference B was replaced by a filler item, and in the other condition (*reference B only*; Figure 3.24b) the distractor was replaced by a filler. The condition with both reference B and the distractor present was labeled *full pair B* (Figure 3.24a). A baseline condition (*pair A only*; Figure 3.24d) consisted of the same trials but with both items of pair B replaced by fillers.

Expectations in this experiment were inspired by the model described in Chapter 2, in particular Section 2.3.3, where it has been demonstrated how different grounding scenarios, defined by combinations of spatial phrase and visual scene, lead to grounding processes of varying complexity. It was hypothesized on this basis that the presence of items forming a relational pair would lead to more complex processes than items not forming a relational pair, with each non-filler item in the display potentially being brought into the attentional foreground multiple times.

It was therefore hypothesized that a relational pair composed of distractor and reference B would evoke attraction stronger than the sum of effects seen in conditions where either only a distractor was added to the visual display, or only another reference item. In other words, an effect based specifically on the presence of pair B was expected to manifest as an interaction between the two factors of reference B presence and distractor presence.

For the case of the competing hypothesis being true, that is, if a stereotypical class of mechanisms was responsible for the observed attraction, it was expected that the individual attraction caused by each of multiple items would combine additively. A distractor item, for instance, would contribute a specific amount of attraction, as would an additional item in reference color, and if both were present, the sum of the two individual contributions would be observed. Therefore, no interaction was expected to occur in this case.

In summary, experiment six compared the combined but relation-independent impact of a distractor and a reference item to their relation-based impact through examining whether the presence of one item moderated the impact of the other. If so, it could be concluded that attraction effects were based on flexible processes of spatial language grounding rather than stereotypical processes applied to isolated items.

An accessory hypothesis tested in this experiment was that attraction toward an item in reference color would occur even if that item was not particularly close to an item sharing the target color. This was examined by assessing the impact of reference B within the condition *reference B only*.

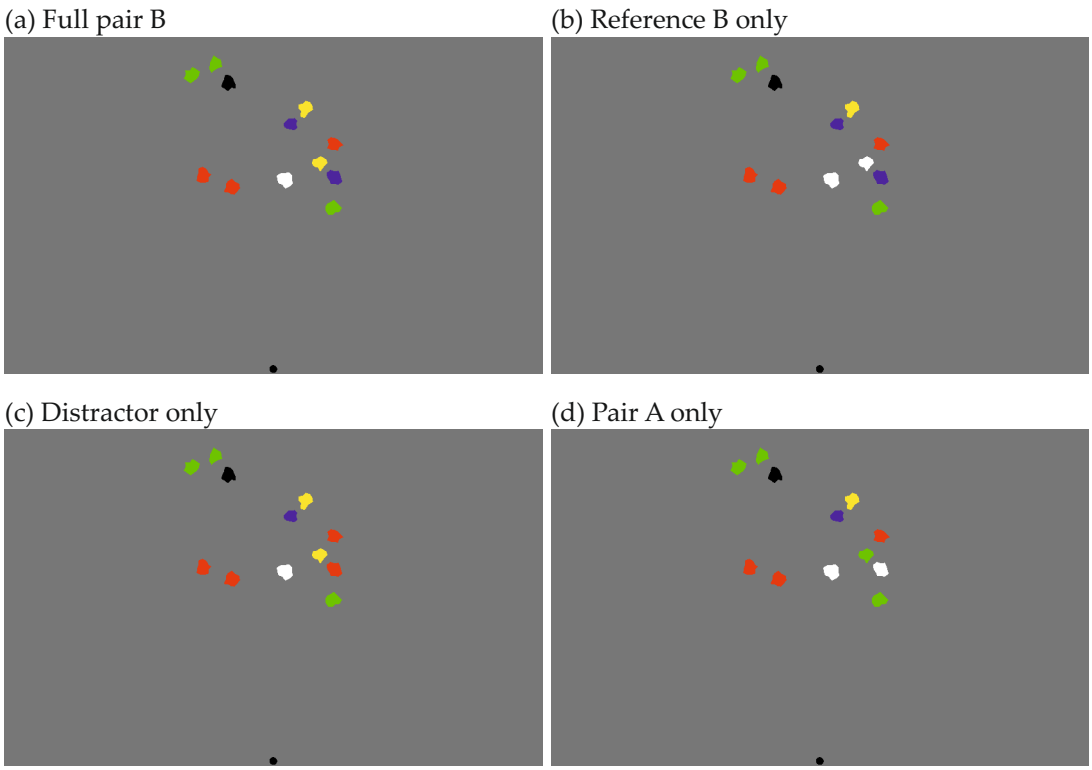


Figure 3.24: Example displays for each of the four new conditions in experiment six for the spatial phrase "The yellow one to the right of the blue one".

3.6.1 Methods

Participants

Twenty participants (ten female, ten male) with a mean age of 23.3 years (SD = 3.4 years) were recruited by notices around the local campus, signed informed consent (Appendix B), and received €15 for participation. All participants were right-handed, as determined by the Edinburgh Handedness Inventory (Oldfield, 1971). The participants were naïve to the experimental hypotheses, native German speakers (one 27 year old female participant had learned German at the age of twelve), had self-reported normal or corrected-to-normal vision, and no color vision deficiencies (Appendix A).

Procedure

The general procedure differed from that of experiment one (see *Procedure*, p.69) only in that trials with incorrect responses were appended at the end of the trial list in order to be repeated later (i.e., trials in which the selected item was not the one best-matching the spatial term). However, a trial would be presented two times at most. This change was made to increase the amount of usable data, under the assumption that most incorrect responses were based

on random errors due to fatigue and similar temporary effects.

The general display configuration was the same as in experiment one, that is, participants responded by moving vertically, with the start marker in the bottom position and the stimulus region centered horizontally in the upper portion of the screen.

Material

Generating visual displays and facilitating balancing The general procedure for creating visual displays was similar to previous experiments. Main differences included that balancing of trial numbers over balancing categories was facilitated already during display generation, and that a second relational pair was added to each display instead of a simple distractor item. The new aspects and differences to the previous experiments are detailed in the following.

The first step of display generation was to construct multiple different spatial configurations of a reference and a target item (pair A) for each spatial term (Figure 3.25b shows the possible target placements for 'left of'). The fit cut-off defining the target item region (Figure 3.25a) was increased to 0.7 (opposed to 0.6 in previous experiments) in order to reduce the maximum spatial extent of pair A, which made later placement of pair B easier. Minimum item border distance was unchanged (0.5 mm/0.04°). The result was a pool of 38 pair A configurations for each of the four spatial terms (with the four sets differing only by rotation around the reference location), equaling a total of 152 configurations.

Next, a second relational pair (pair B) was placed in these displays. Pair B was obtained by mirroring pair A from the respective display either horizontally for horizontal axis spatial terms or vertically for vertical axis spatial terms (item shapes were created independently for pair A and pair B). Thus, the relation instantiated by pair B was the opposite of that given in the spatial phrase and instantiated by pair A. Pair B consisted of reference B, which shared the color of reference A, and the distractor item, which shared the color of the target item. Pair B could be placed either entirely (i.e., both of its items) on the left side of the direct path to the target, or entirely on its right side. It was not allowed to be placed on or overlap the direct path.

The 152 configurations created so far were used as a pool from which configurations were taken to create displays that realized the different possible combinations of pair B side, spatial term, and on-screen target position. In short, creating a specific combination of these included picking a configura-

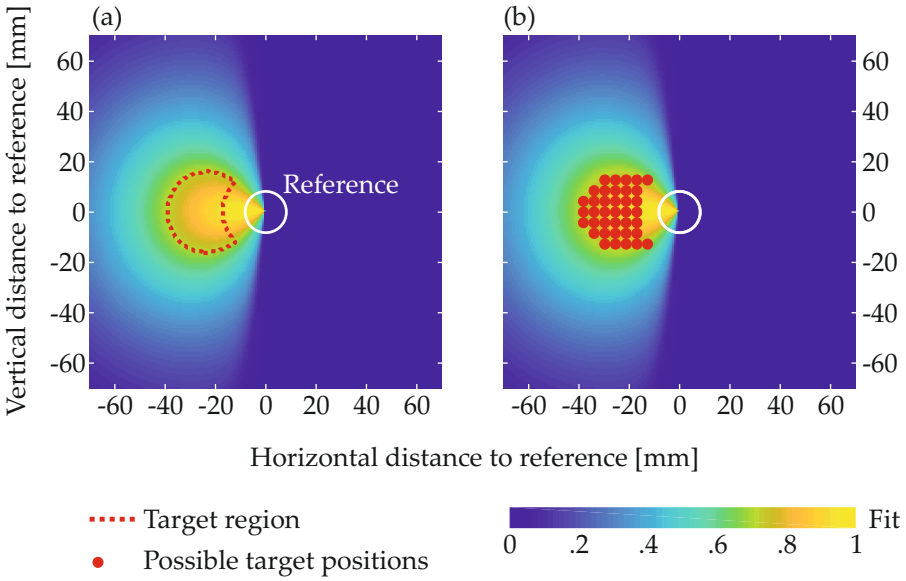


Figure 3.25: Item placement in experiments six and seven based on fit constraints for the spatial template for 'left of'. (a) The region generally eligible for target placement defined by fit cut-off and minimal item distance (target center must lie within this region). (b) All possible target positions used in the experiment.

tion from the 38 ones instantiating the desired spatial term, translating that configuration to place the target in the desired on-screen target position, and then placing pair B on the desired side of the direct path.

To understand the rationale according to which pair B was placed either to the left or to the right of the direct path, it is important to know that during analysis effects of pair B were estimated from the response data in the same way effects were extracted in the previous experiments. As described in *Balancing the effects of potentially confounding items (p.81)*, this involved composing balanced overall means from multiple conditions to equalize the impact of trials from the four balancing categories. In the current experiment, the 'item' of interest was pair B and the potentially confounding items were reference A and the CoM. Pair B was treated as a single item for matters of balancing, which was warranted by the constraint that it was always entirely to the left or to the right of the direct path. The relevant balancing categories are listed in Table 3.10.

The scheme by which balanced overall means were formed to isolate the effect of pair B is shown in Figure 3.26. Each panel in this figure corresponds to a specific combination of pair B side, spatial term, on-screen target position, and balancing category. Note that the balancing scheme is analogous to that of the distractor effect (Figure 3.8) in the previous experiments. As the missing outline colors in Figure 3.26 indicate, overall means were computed only across spatial terms and not separately for each spatial term axis because

Table 3.10: Balancing category labels for the effect of pair B. 'Side' refers to the side of the respective item relative to the direct path in comparison to the side of pair B relative to the direct path.

	CoM side	
Reference A side	Same	Different
Same	rs/cs	rs/cd
Different	rd/cs	rd/cd

spatial term effects were expected to be canceled out through balancing, removing the need to consider spatial term axes separately.

In contrast to the previous experiments, balancing was facilitated already during display generation by placing pair B on either the left or the right side of the direct path such that the required trial numbers were achieved for each panel of Figure 3.26. This was possible since the fit constraints governing the placement of pair B were less restrictive as those governing distractor placement in previous paradigms.

Concretely, for each panel in Figure 3.26, the list of the 38 pair A configurations instantiating the correct spatial term was cycled through, and for each case therein, the configuration was translated to the on-screen target position prescribed by the panel, after which pair B was placed on the side required to satisfy the balancing category prescribed by the panel. This was done until a pre-specified number of displays was obtained for each of the panels. Configurations from the set at hand were reused if more than 38 displays were desired (with pair B placed at a different position each time this happened).

Figure 3.26 (on following page): Schematic depiction of possible item configurations in experiments six and seven for each combination of pair B side, on-screen target position, and spatial term. Balancing categories are indicated in the bottom right of each panel. Panels are arranged by pair B side to illustrate which conditions were combined and compared to examine the effect of pair B. Overall means for statistical comparisons were computed over sets of panels sharing outline style. Target (T) and distractor (D) are shown as green dots, reference A (RA) is shown in red, as well as reference B (RB), and the CoM (fillers not shown) is depicted by a black diamond. The start marker is shown as a black dot and light gray crosses indicate possible on-screen target positions.

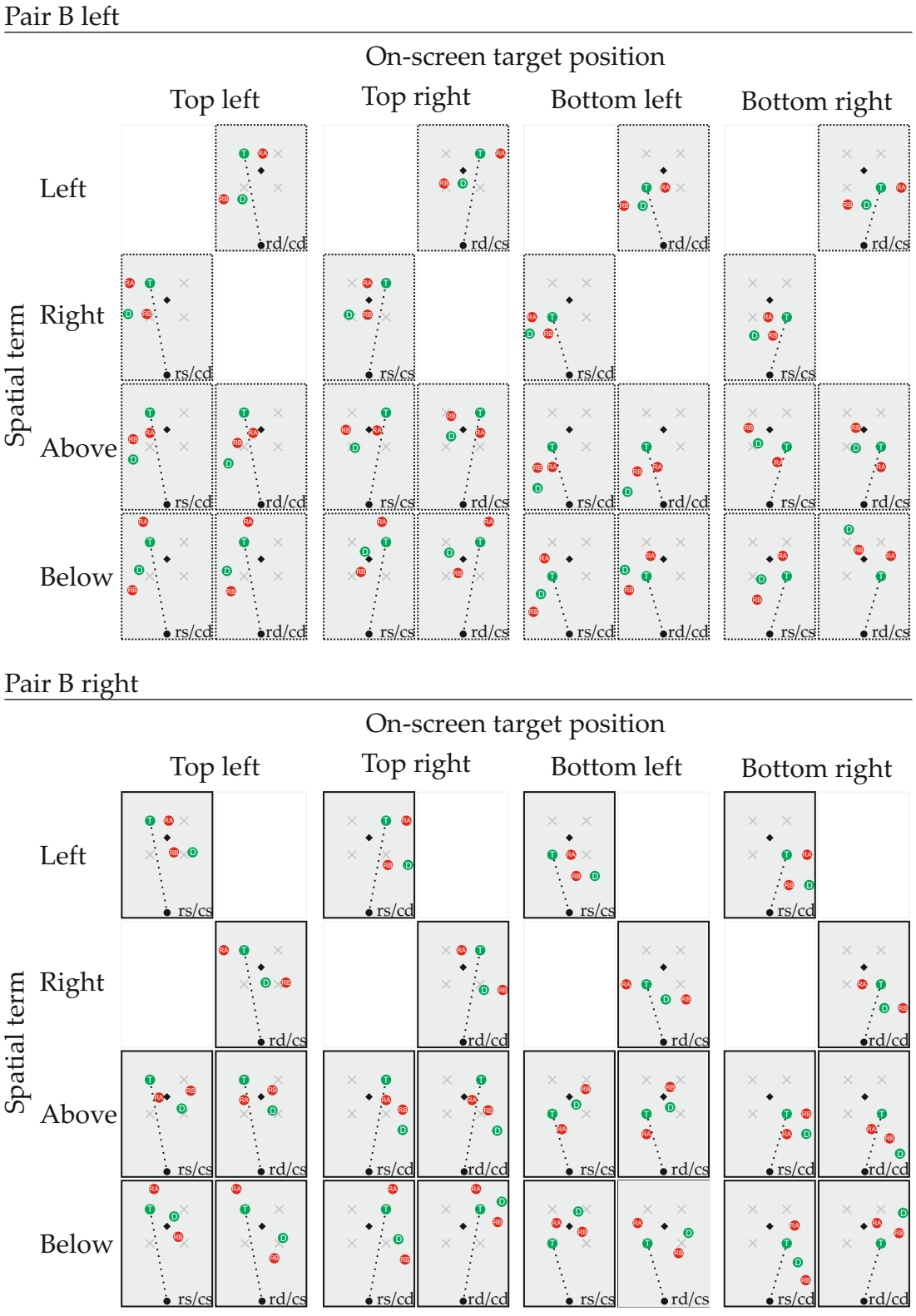


Figure 3.26: Caption on previous page.

The number of displays created for each panel was either 80, for horizontal axis spatial terms, or 40, for vertical axis spatial terms. The difference in these numbers accounts for the fact that, as in the previous paradigms, conditions with vertical axis spatial terms allowed realizing two balancing categories, while those with horizontal axis spatial terms allowed only one. By creating double the number of displays for the conditions compatible with only one balancing category it was ensured that each overall mean included the same number of trials from each spatial term.

As mentioned above, configurations of reference A and target item were used multiple times for the same panel when the number of desired trials per panel exceeded the number of different pair A configurations. Apart from that, configurations were also reused due to the issue that some conjunctions of a pair A configuration and an on-screen target position did not allow to realize certain balancing categories. This occurred when, after translating the configuration to place its target item in the on-screen position, reference A happened to lie on the same or different side as the CoM without this relative side fitting the demand of the balancing category at hand.

For instance, when reference A and CoM were on different sides of the direct path, the balancing category rs/cs could not be realized. This is akin to the issue described in *Balancing the effects of potentially confounding items (p.81)* that specific conjunctions of horizontal axis spatial terms and on-screen target position precluded some balancing categories because the side of the reference was prescribed by the spatial term. Since for horizontal terms this was solved by altogether omitting the problematic combinations (empty panels in Figure 3.8, 3.9, and 3.26), the issue was relevant only for vertical axis spatial terms in the current context. For these, the solution was to not use those pair A configurations that could not satisfy the balancing category of the panel at hand and instead use the remaining configurations more often. As a result, only 15 or 23 of the 38 configurations were used for each panel in Figure 3.26 with a vertical axis spatial term. Whether 15 or 23 were used depended on the conjunction of spatial term and on-screen target position, as this determined the slope of the direct path (whether it leaned left or right) and thus for how many configurations reference A ended up to the left or to the right of it.

Note that the sets of configurations used to realize a particular balancing category complemented each other to the full set of 38 configurations when pooled across the two vertical axis spatial terms. For instance, for the combination of 'above' with the top left on-screen target positions and pair B on the right side (see Figure 3.26), rs/cs could be satisfied by only 15 configu-

rations, since for 'above' on average across configurations reference A was located straight below the target item, as can be derived from Figure 3.25b. Conversely, for the same combination and 'below' the other 23 configurations were used, since in this case reference A tended to be located straight above the target, meaning it was more often situated to the right of the direct path and thus on the same side as the CoM.

Apart from the side of placement, the regions in which pair B was allowed to be placed were restricted by the following constraints. All items had to lie entirely within the stimulus region, their outer radiuses not overlapping the region's border (the stimulus region is illustrated in Figure 3.2). The minimum distance of item borders to each other was the same as in previous experiments (0.5 mm/0.04° v.a.). Moreover, the outer radiuses of the items of pair B had to be separated from the direct path by at least 3 mm (2.45° v.a). This minimum distance was meant to ensure that deviation toward the pair would be visible as deviation from the direct path. Conversely, the centers of the items of pair B were never placed further from the direct path than 56.8 mm (4.64° v.a), in order to place pair B in the vicinity of the actual on-screen target positions and prevent participants from ruling out pair B as a potentially relevant candidate for processing solely due to its remoteness from the usual target placements.

Finally, there were two constraints that prevented interference between the items of pair A and pair B in terms of how well their items matched the spatial term. First, pair B had to be placed such that the distractor item's position in relation to reference A provided a worse match for the spatial term than the position of the target item (as otherwise the two items would have switched roles), with a minimum fit difference of 0.25. Second, the target item's position in relation to reference B had to provide a worse match for the trial's spatial term than its position in relation to reference A, where again the difference in fit values had to be higher than 0.25.

This set of constraints defined a region eligible for the placement of pair B for each combination of on-screen target position and pair A configuration. An example for this is shown in Figure 3.27, where panel a shows the placement template for placement on the left side of the direct path, and panel b shows the region for placement on the right side. Note that the yellow area marks locations at which reference B could be placed, but that the extent of that region also takes into account the constraints for distractor placement, given the relational pair at hand. The placement of pair B within the eligible region was random. Given the large number of trials, this lead to a relatively

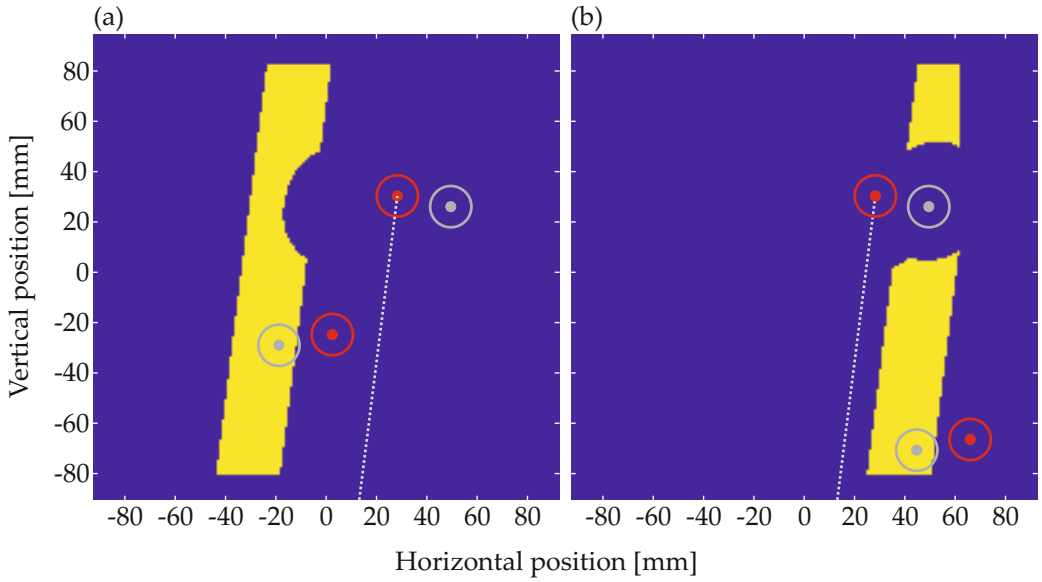


Figure 3.27: Example for a template of possible positions for pair B within the stimulus region (for spatial term 'left of'). Yellow regions represent areas where reference B could be placed without violating any constraints pertaining to the distractor item or reference B itself, given the current relational pair. The upper two circles in each panel represent the outer border of the target (red) and reference A (gray). The lower two circles represent the distractor (gray) and reference B (red), whose position in the figure represents one possible placement. Panel a shows the template for the case where pair B had to be placed to the left of the direct path, panel b shows the opposite case. The dotted gray line is the direct path.

uniform coverage of the available space.

To summarize display creation thus far, there were 32 conditions, that is, combinations of pair B side, spatial term, and on-screen target position. Half of these 32 conditions allowed to realize one balancing category, and for each of these 80 displays were created. The other half allowed realizing two balancing categories, and each of these was represented by 40 displays. This amounts to a total of 2560 visual displays composed of four items each.

The basic displays were finalized by adding eight filler items to each, governed by the same constraints as in previous experiments. No opposite distractor was used this time. The 2560 displays were then equally distributed onto the 20 participants so that each was assigned 128 visual displays. The assignment was random except for ensuring that the overall ratio of trial numbers in the different conditions and balancing categories was retained within each participant, to preserve balancing on the participant level.

In a final step of trial creation, the displays were modified to realize the new conditions pertaining to the presence versus absence of reference B and the distractor, respectively (Figure 3.24). This was done separately based on the already assigned trial set of each participant. For each of the four new conditions the respective participant's set of 128 trials was reused, such that

each participant had to complete 512 trials in total. The condition *full pair B* was represented by the unmodified displays. For the condition *distractor only*, the color of reference B was changed randomly to one of the filler colors, so that only the distractor item remained of pair B. For the condition *reference B only*, the color of the distractor was changed randomly to one of the filler colors, so that only reference B remained of pair B. For the condition *pair A only*, the colors of both the distractor and reference B were changed randomly (and independently) to one of the filler colors, so that only pair A remained.

Spatial phrases Spatial phrases were the same as in experiment one (i.e., only a single form was used).

Analysis

Balancing Overall means for statistical testing were obtained by applying the same method as that used for the previous experiments. The employed scheme is shown in Figure 3.26, where means were computed over the averages of trials from panels with shared outline style. As described in the previous section, only means across all spatial terms were computed (i.e., not for each spatial term axis). Overall means were separately computed within each of the conditions *full pair B*, *distractor only*, *reference B only*, and *pair A only*.

Since for this experiment balancing was facilitated already during the display creation, imbalances removed by this balancing could only occur as a result of incorrect responses (which were rare due to incorrect trials being repeated as described above) and trajectories exceeding curvature threshold.

For considering the effect of reference B in the condition *reference B only* the same balancing scheme was used.

Statistical analysis The focus of the current experiment was to probe for an interaction between the presence of the distractor and reference B. For this, the impact of pair B side in the four conditions *full pair B*, *reference B only*, *distractor only*, and *pair A only* was assessed and compared. These conditions are based on the factors *reference B presence* and *distractor presence*, each with the levels *present* and *absent*. Table 3.11 summarizes factors and conditions of this 2×2 within-subjects design.

Difference scores between pair B right and pair B left at each time step were computed within each of the four conditions, and the resulting time-series of difference scores were compared between conditions. Cell means

Table 3.11: Conditions in the 2 × 2 within-subjects design of experiments six (and seven).

	Distractor	
Reference B	Present	Absent
Present	Full pair B	Reference B only
Absent	Distractor only	Pair A only

at each trajectory time step were subjected to a two-way repeated measures ANOVA (resulting in 151 ANOVAs). The employed alpha level was $p < 0.04$ to retain an overall alpha level of $p < .05$, since apart from the ANOVAs one additional planned comparison was conducted with $p < .01$ (see below). Effect sizes for the ANOVAs were computed as partial eta-squared, η_p^2.

A single planned comparison using paired t-tests with $p < .01$ was conducted to assess the effect of reference B side in the condition *reference B only* (comparing reference B right to reference B left of the direct path).

As before, bootstrapping was performed for each comparison, obtaining a separate criterion for each main effect, interaction, and for the single planned comparison, each bootstrap being based on $p < .01$ for the overall criterion. Where data was tested using ANOVAs, bootstrap samples were subjected to equivalent ANOVAs.

Overall movement times over trials with the item of interest left and right, respectively, were computed within each of the four conditions.

3.6.2 Results

A total of 10240 trajectories was obtained. Of these, 9554 (93.3 %) were below curvature threshold (M = 477.7, SD = 18.36 equaling M = 93.3 %, SD = 3.59 %).[5] Of the non-curved trajectories, 98.17 percent (9379) were correct responses and thus entered further analysis (91.59 % of all obtained trajectories). Movement onset was generally close to the center of the start marker (M = 1.72 mm, SD = 1.84 mm).

Participants achieved a mean accuracy of 98.17 percent (SD = 1.89 %) and their mean movement time was 1101 milliseconds (SD = 121 ms). Note that the numbers reported so far were not affected by balancing but based on simple averaging over the respective trial ensembles; mean data reported from here on was obtained according to the balancing scheme described above.

[5]Note that, over participants, an average of 42.5 (SD = 19.33) trials were presented twice due to incorrect responses in the first presentation, as described earlier. If the second presentation of a trial was responded to correctly, only this response was included in the remaining analyses.

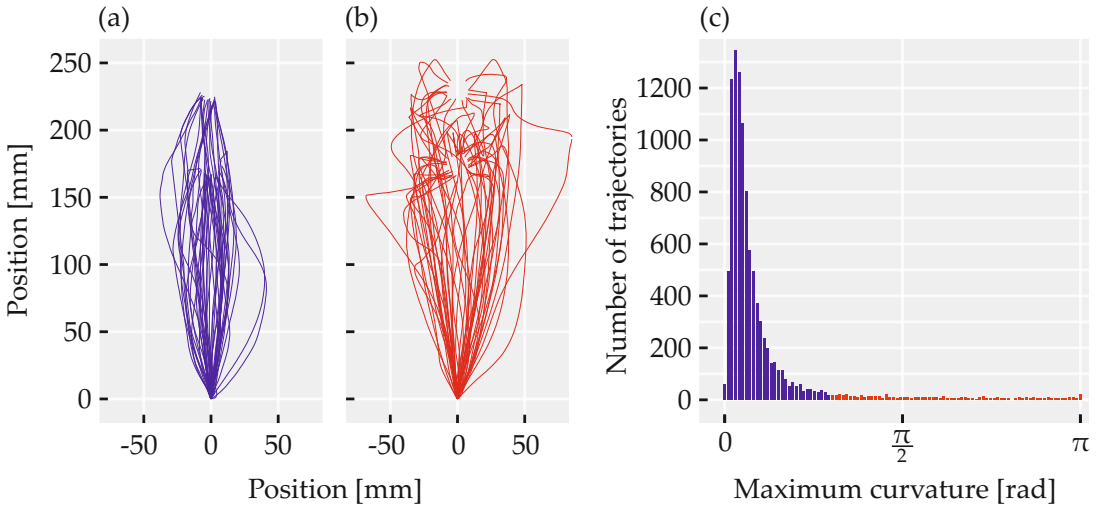


Figure 3.28: Fifty randomly selected trajectories obtained in experiment six that were (a) below curvature threshold and (b) exceeding curvature threshold. Panel c shows the overall distribution of trajectories over maximum curvature values, where red bars correspond to trajectories that were discarded due to high curvature. Only correct responses are shown.

Condition-specific movement times are listed in Table 3.12, with an apparent tendency of movement duration to increase in the order: *pair A only* (lowest), *reference B only*, *distractor only*, and *full pair B* (highest).

Figure 3.28 shows fifty examples for trajectories below and above curvature threshold (panels a and b, respectively) along with the empirical distribution over maximum curvature value for all correct responses (panel c), with red bars indicating curvature above threshold (i.e., trajectories excluded from other analyses). For the distribution Hartigan's dip test indicated no bimodality ($p > .05$).

To provide a sense of the deviation toward items of interest in this paradigm, Figure 3.29 shows mean trajectories for each condition and for each side of the item of interest (pair B as a whole, the distractor, or reference B).

The comparison of left and right mean trajectories within condition *ref-*

Table 3.12: Movement times and standard deviations (SD) for each condition in experiment six.

	Item of interest side					
	Left		Right		Overall	
Condition	Mean	SD	Mean	SD	Mean	SD
Full pair B	1130	124	1161	125	1145	121
Distractor only	1112	128	1114	125	1113	125
Reference B only	1076	125	1096	125	1086	124
Pair A only	1061	123	1073	115	1067	118

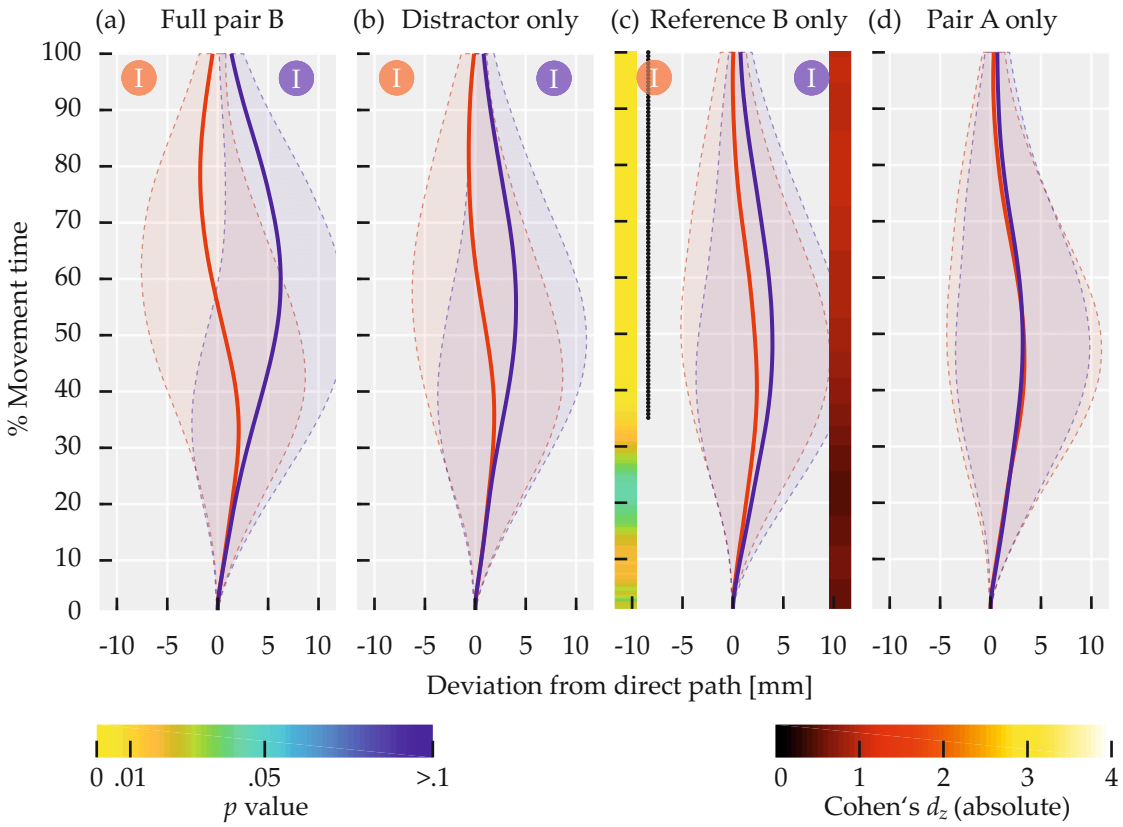


Figure 3.29: Deviation from the direct path in experiment six. Shown are mean trajectories for each condition and for each side of the item of interest (pair B as a whole, the distractor, or reference B), indicated by red and blue circles labeled 'I' for correspondingly colored trajectories. The image map on the left side of panel c indicates the p values from the t-tests at the respective time steps, the one on the right side indicates effect sizes (absolute Cohen's d_z), and the black dotted line on the left side spans time steps with significant differences between the mean trajectories. Transparent regions delimited by dashed lines indicate between-subjects standard deviation.

erence B only (Figure 3.29c) was significant, and indicated an extensive bias toward reference B with an early onset. It spanned 99 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 4 time steps. The sequence of significant differences extended from 35.1 to 100 percent of movement time, with the minimum p value occurring at 78.81 percent movement time ($t(19) = 4.62$, $p < .001$, $d_z = 1.03$).

Results of the repeated measures ANOVAs of trajectory divergence in each condition (i.e., with the factors distractor presence and reference B presence) are shown in Figure 3.30. The main effect of distractor presence (Figure 3.30a) was significant at 96 time steps, the sequence extending from 37.09 to 100 percent of movement time, with the minimum p value occurring at 72.19 percent movement time ($F(1, 19) = 111.69$, $p < .001$, $\eta_p^2 = 0.855$). However, the sequence length did not reach the criterion obtained from the bootstrap (based on overall $p < .01$) of 100 steps (which was likely due to the

bootstrap method being prone to yielding overly conservative criteria in the case of particularly strong effects as observed here, as described in *Bootstrapping*, p.91).

The main effect of reference B presence (Figure 3.30b) was significant at 109 time steps, the sequence extending from 28.48 to 100 percent of movement time, with the minimum p value occurring at 80.79 percent movement time ($F(1,19) = 24.9564$, $p < .001$, $\eta_p^2 = 0.568$). This effect did exceed the bootstrap criterion for overall significance (based on overall $p < .01$) of 101 time steps.

The interaction between reference B presence and distractor presence (Figure 3.30c) was significant as well, spanning 29 time steps, the sequence extending from 72.85 to 91.39 percent of movement time. The interaction exceeded the bootstrap criterion for overall significance (based on overall $p < .01$) of 13 time steps. The minimum p value occurred at 85.43 percent movement time ($F(1,19) = 5.669$, $p = .0279$, $\eta_p^2 = 0.230$); an interaction plot for this time step is shown in Figure 3.30d, illustrating more clearly the stronger impact of adding one item to the display when the other is as well present, compared to adding the same item to a display that does not contain the other item.

3.6.3 Brief discussion

The most important result of experiment six was the over-additive interaction between the presence of the distractor item and the presence of the additional reference item. The effect caused by the joint presence of an additional reference item and a distractor item was larger than what would have been expected based on the impact of either item in isolation. This suggests that, during grounding, additional processes take place when more items are present that may instantiate the relational phrase being grounded.

That an interaction was observed also strengthens the notion that attraction effects in the experiments so far were the product of organized and flexible processes of spatial language grounding — whose complexity increase when more potential role fillers are present (which is also in line with the model described in Chapter 2).

In addition to the interaction, the main effects of reference and distractor presence were significant and strong; this is in line with the attraction seen in previous experiments. However, the presence of the interaction makes further interpretation of the main effects difficult.

The accessory test for an effect of the additional item in reference color (i.e.,

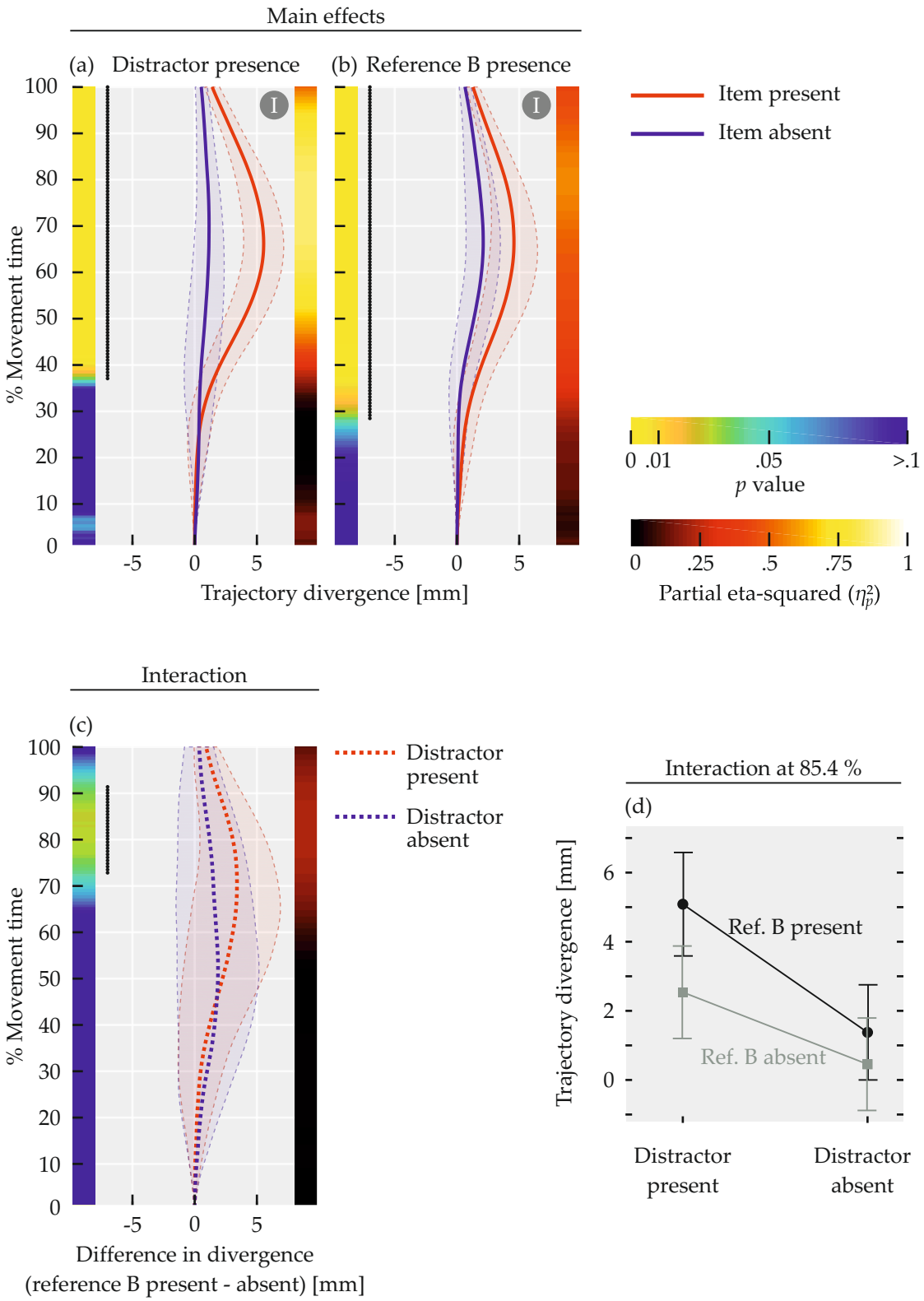


Figure 3.30: Result of the ANOVAs performed on trajectory divergence scores from experiment six. Panel a illustrates the main effect of distractor presence, panel b shows the main effect of reference B presence (for these panels, circles labeled 'I' indicate the effective side of the item of interest, that is, distractor or reference B). Panel c illustrates the interaction of these factors as the impact of reference B presence on trajectory divergence when the distractor is present or absent. Panel d shows a standard interaction plot for the point of movement time at which the lowest p was observed. Transparent regions delimited by dashed lines, and error bars in (d), indicate standard deviation between participant means. Note that x-axis scaling differs from plots that show mean trajectories.

reference B) showed that it lead to attraction, even though there was no item in target color in its immediate vicinity. The effect had a somewhat earlier onset than the reference effects seen in the other experiments, with a trend toward significance even in early portions; the origin of this is unclear. Interestingly, the effect was close in effect size and mean difference to the reference effects seen in previous experiments even though the additional reference item was placed more freely than the veridical reference item and thus tended to be more remote from the target and from the direct path. This might hint that participants did not move directly toward the additional reference but were gradually attracted toward its location.

Movement times appeared to increase in the order: *pair A only*, *reference B only*, *distractor only*, and *full pair B*. This seems plausible given that more pronounced attraction likely leads to longer mouse paths on average, but there were no formal assessments in this regard.

Finally, the higher accuracy in the current experiment was based on presenting trials for which an incorrect response was given in the first presentation again at a later point (but two times at most). The very high accuracy achieved with this method suggests that many incorrect responses in previous experiments were not due to participants judging goodness of fit differently than expected, but arose from momentary factors. This again confirms that spatial template shapes captured the semantics of spatial terms to an adequate degree for the current purpose.

In summary, experiment six suggests that relational pairs that represent potential referents for a spatial phrase lead to additional grounding processes in sensorimotor representations compared to isolated additional items in target or reference color. Moreover, an additional item in reference color leads to attraction similar to the reference item of the relational pair referred to in the spatial phrase.

3.7 Experiment seven: Attraction toward multiple items

The results of experiment six were interpreted as reflecting additional processes during grounding evoked by a pair of items posing a potential referent of the spatial phrase. An alternative is that any two potentially task-relevant items on the same side of the direct path may interact to cause a degree of attraction that transcends the sum of the items' individual effects. This could

not be ruled out based on experiment six since it did not include a condition where two additional items were present on the same side of the direct path without forming a relational pair.

This was probed in the current experiment. The only difference to experiment six was that in the conditions *reference B only* and *distractor only* the removed pair B item was replaced not by a filler color, but by the color of the remaining pair B item. That is, in the condition *reference B only* both items shared the color of reference A (Figure 3.31b), and in the condition *distractor only* both items shared the color of the target item (Figure 3.31c).

Hypotheses were based on the assumption that the interaction in experiment six was indeed the result of additional grounding processes. It was thus hypothesized that trajectory divergence between pair B sides would be stronger in the condition *full pair B* than in the condition *distractor only* (two distractor items) and stronger than in the condition *reference B only* (two reference B items).

In addition to this, the attraction effect exerted by an additional relational pair was directly tested in this experiment by comparing mean trajectories between pair B sides within the condition *full pair B*. Further comparisons tested whether the presence of two items lead to effects comparable to or larger than those exerted by single items, by comparing mean trajectories in the conditions *reference B only* and *distractor only*. This was done to probe to which degree effects were comparable to previous experiments, and to ensure that doubled items did not allow identifying the two irrelevant items in a preattentive manner based on the larger amount of a single task-relevant color within a small area of the visual display (which was assumed to manifest as strongly decreased attraction effects; e.g., Song & Nakayama, 2006, show a similar effect of perceptual grouping, but based on a much larger number of uniformly colored items).

3.7.1 Methods

Participants

Twenty participants (11 female, nine male) with a mean age of 27.1 years (SD = 7.7 years) were recruited by notices around the local campus, signed informed consent (Appendix B), and received €15 for participation. All participants were right-handed, as determined by the Edinburgh Handedness Inventory (Oldfield, 1971). The participants were naïve to the experimental hypotheses, native German speakers, had self-reported normal or corrected-

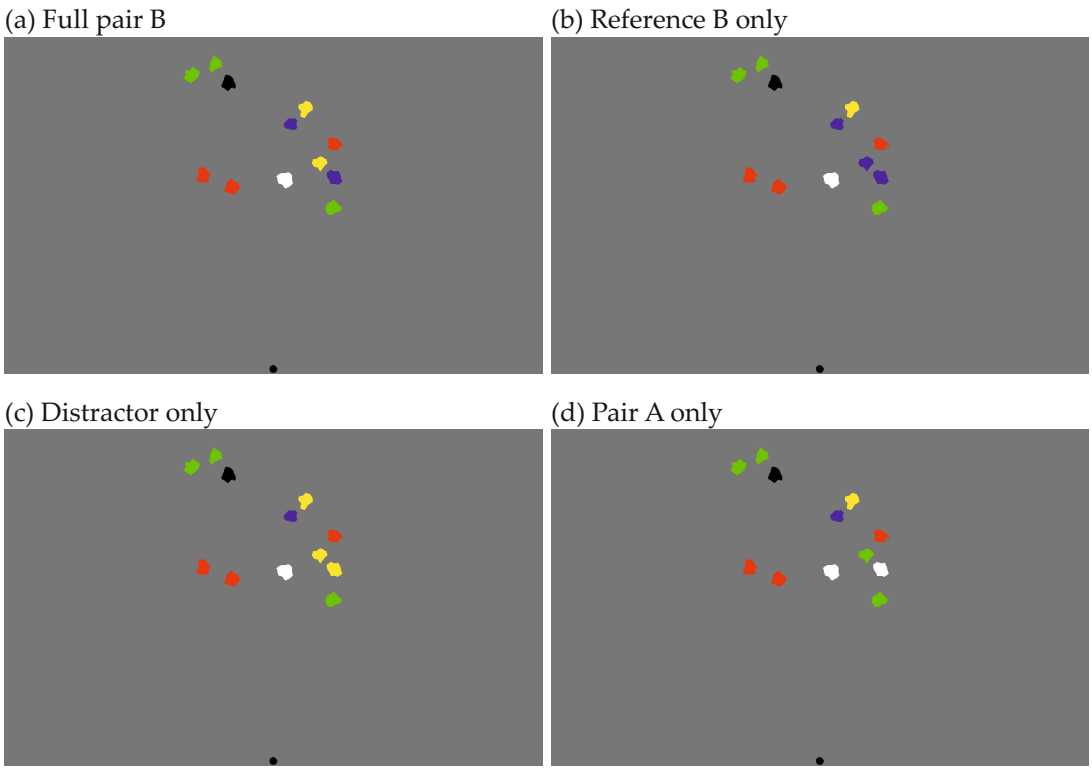


Figure 3.31: Example displays for each of the four conditions in experiment seven for the spatial phrase "The yellow one to the right of the blue one".

to-normal vision, and no color vision deficiencies (Appendix A).

Procedure

The procedure was identical to experiment six.

Material

All used materials were the same as in experiment six, with the exception that 'absent' pair B items now were colored in the same color as the remaining pair B item rather than being turned into a filler. This means that visual displays were fully identical to experiment six in the conditions *full pair B* and *pair A only*. In the condition *distractor only*, the former reference B was given the same color as the distractor, resulting in the presence of two distractor items. In the condition *reference B only*, the former distractor item was given the same color as the reference B item, resulting in the presence of two additional items with the same color as the reference item.

Analysis

Balancing Balancing followed the same logic as in experiment six (see Section 3.6.1 and Figure 3.26). The balancing scheme described there was used for all comparisons in experiment seven.

Statistical analysis Mean trajectories over trials where pair B or the replacing items were to the left of the direct path were compared to those where they were to the right of the direct path. This was done within each of the conditions *full pair B*, *distractor only*, and *reference B only*.

Difference scores between pair B right and pair B left at each time step were computed within each of the four conditions *full pair B*, *distractor only*, *reference B only*, and *pair A only*. The resulting time-series of difference scores were compared between conditions for the pairings *full pair B* versus *distractor only* and *full pair B* versus *reference B only*.

Each of the five comparisons used paired t-tests with $p < .01$ and bootstrap confirmation also using $p < .01$. Movement times over trials with the item of interest left and right, respectively, were computed within each of the four conditions.

3.7.2 Results

A total of 10240 trajectories was obtained. Of these, 9529 (93.06 %) were below curvature threshold (M = 476.45, SD = 18.63 equaling M = 93.06 %, SD = 3.64 %).[6] Of the non-curved trajectories, 98.79 percent (9414) were correct responses and thus entered further analysis (91.93 % of all obtained trajectories). Movement onset was generally registered close to the center of the start marker (M = 1.84 mm, SD = 3.12 mm).

Participants achieved a mean accuracy of 98.77 percent (SD = 2.61 %) and their mean movement time across conditions was 1074 milliseconds (SD = 125 ms). Note that the numbers reported so far were not affected by balancing but based on simple averaging over the respective trial ensembles; mean data reported from here on was obtained according to the balancing scheme described above.

Condition-specific movement times are listed in Table 3.13. Similar to experiment six, there appeared to be a tendency toward higher movement time

[6] Note that an average of 39.05 (SD = 21.48) trials per participant were presented twice due to incorrect responses in the first presentation, as described earlier. If the second presentation of a trial was responded to correctly, only this response was included in the remaining analyses.

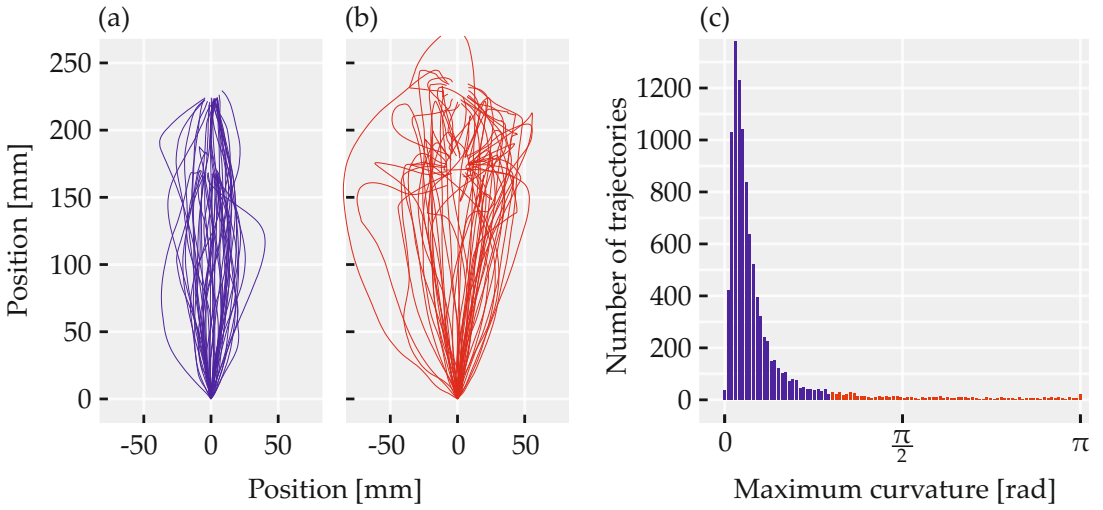


Figure 3.32: Fifty randomly selected trajectories obtained in experiment seven that were (a) below curvature threshold and (b) exceeding curvature threshold. Panel c shows the overall distribution of trajectories over maximum curvature values, where red bars correspond to trajectories that were discarded due to high curvature. Only correct responses are shown.

in the order: *pair A only* (lowest), *reference B only*, *distractor only*, and *full pair B* (highest).

Figure 3.32 shows fifty examples for trajectories below and above curvature threshold (panels a and b, respectively) along with the empirical distribution over maximum curvature value for all correct responses (panel c), with red bars indicating curvature above threshold (i.e., trajectories excluded from other analyses). For the distribution Hartigan's dip test indicated no bimodality ($p > .05$).

Figure 3.33 shows comparisons of mean trajectories between sides of the item of interest (pair B as a whole, the distractor, or reference B) within different conditions. Red and blue circles labeled 'I' in the top of each panel indicate the side of item of interest for the correspondingly colored mean trajectory.

Table 3.13: Movement times and standard deviations (SD) for each condition in experiment seven.

	Item of interest side					
	Left		Right		Overall	
Condition	Mean	SD	Mean	SD	Mean	SD
Full pair B	1111	138	1123	125	1117	130
Distractor only	1098	133	1107	132	1103	131
Reference B only	1046	112	1049	135	1047	122
Pair A only	1043	121	1041	118	1042	119

In *full pair B*, a sustained bias toward pair B occurred (Figure 3.33a), with 105 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 78 time steps. The sequence of significant differences extended from 31.13 to 100 percent of movement time, with the minimum p value occurring at 76.82 percent movement time ($t(19) = 17.24$, $p < .001$, $d_z = 3.86$).

A bias toward the two distractor items in *distractor only* was similarly sustained (Figure 3.33b), with 102 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 35 time steps. The sequence of significant differences extended from 33.11 to 100 percent of movement time, with the minimum p value occurring at 76.16 percent movement time ($t(19) = 11.66$, $p < .001$, $d_z = 2.61$).

A bias toward the two items in reference color in reference B occurred somewhat later (Figure 3.33c), with 73 successive time steps showing significant differences at $p < .01$, exceeding the bootstrap criterion ($p < .01$) of 4 time steps. The sequence of significant differences extended from 52.32 to 100 percent of movement time, with the minimum p value occurring at 92.72 percent movement time ($t(19) = 5.63$, $p < .001$, $d_z = 1.26$).

Figure 3.34a shows trajectory divergence over time (item of interest side right minus left) for each of the four conditions.

Figure 3.34b shows the comparison of the condition *distractor only* (green line) to *full pair B* (red line), revealing that trajectory divergence was larger in *full pair B*. This difference was significant over 66 time steps, exceeding the bootstrap criterion of 32 time steps and extending from 56.95 to 100 percent of movement time. The minimum p value occurred at 73.51 percent movement time ($t(19) = 6.03$, $p < .001$, $d_z = 1.35$).

Figure 3.34c shows the comparison of divergence between *full pair B* (red line) and *reference B only* (blue line). The difference here became significant earlier in movement time and was larger, with 90 time steps showing significant differences, thus exceeding the bootstrap criterion of 88 steps, and extending from 41.06 to 100 percent movement time. For this difference, the minimum p value occurred at 74.17 percent movement time ($t(19) = 10.15$, $p < .001$, $d_z = 2.27$).

3.7.3 Brief discussion

Most importantly, the trajectory divergence caused by an additional relational pair was larger than both trajectory divergence caused by two distrac-

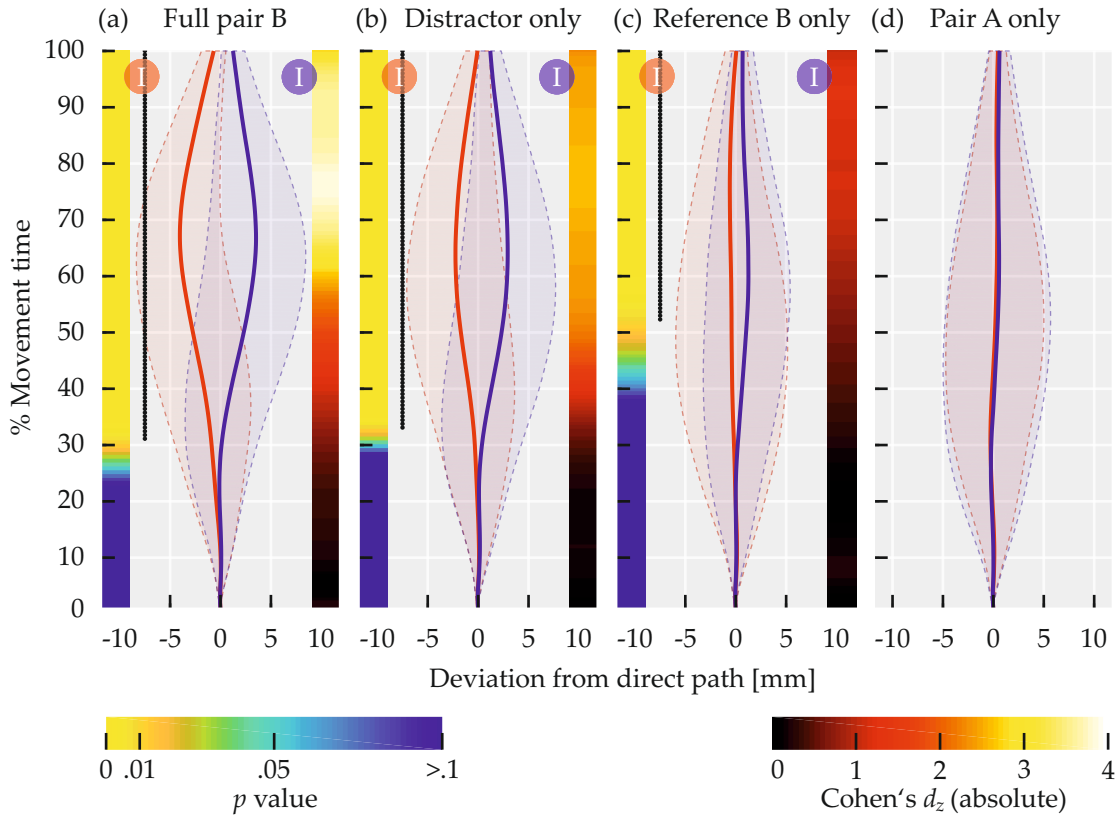


Figure 3.33: Deviation from the direct path in experiment seven. Shown are mean trajectories for each condition and for each side of the item of interest (pair B as a whole, two distractors, or two additional reference items), indicated by red and blue circles labeled 'I' for correspondingly colored trajectories. Image maps on the left sides of panels indicate the p values obtained from t-tests at the respective time steps, those on the right sides indicate effect sizes (absolute Cohen's d_z), and the black dotted lines on the left sides span time steps with significant differences between the mean trajectories. Transparent regions delimited by dashed lines indicate between-subjects standard deviation.

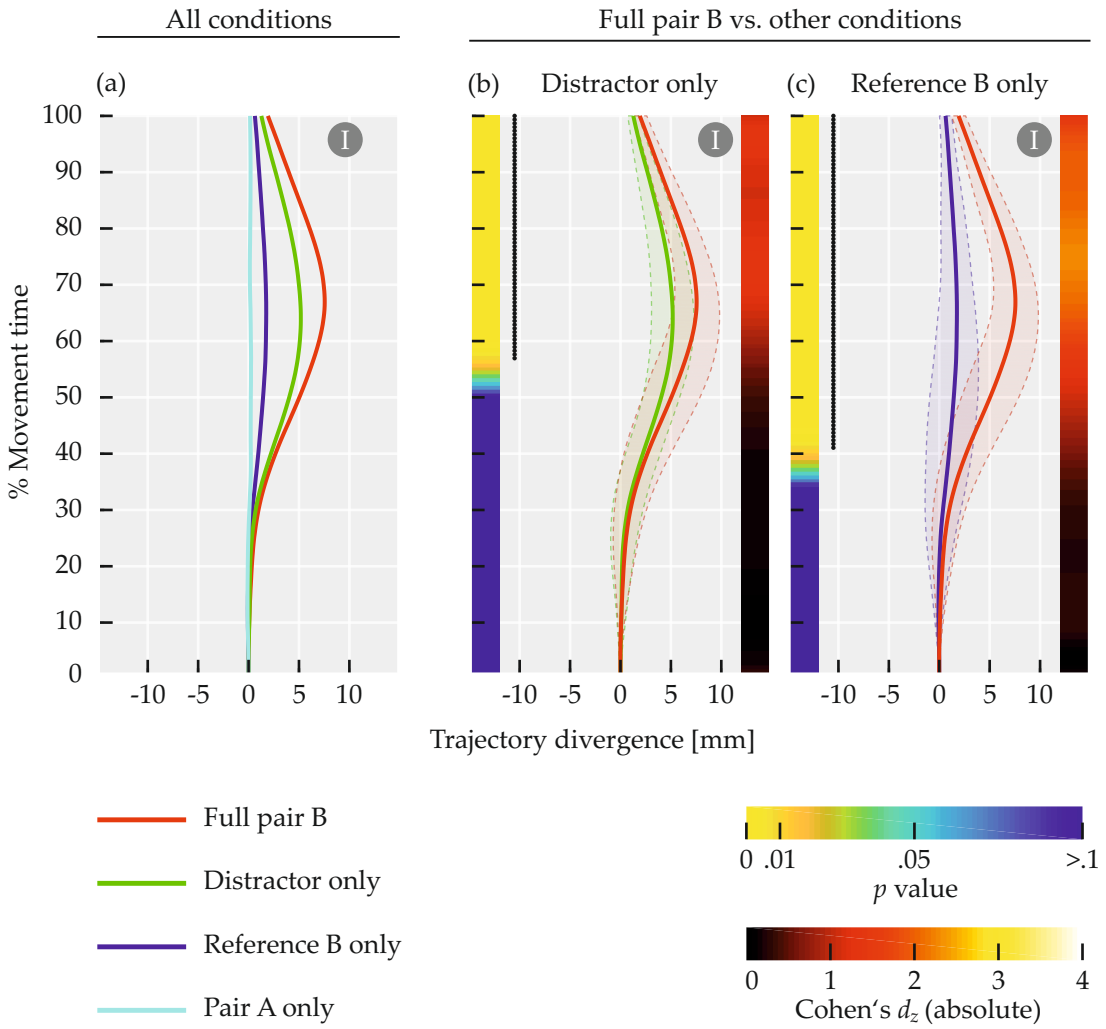


Figure 3.34: Mean trajectory divergence in experiment seven (difference scores between item of interest sides, right minus left, where the item of interest was either pair B as a whole, two distractors, or two additional references). (a) For all conditions in direct comparison. (b) Condition *distractor only* compared against condition *full pair B*. (c) Condition *reference B only* compared against condition *full pair B*. Gray circles labeled 'I' indicate the effective side of the item of interest, meaning that divergence in that direction is equal to a bias toward the item. Transparent regions delimited by dashed lines indicate between-subjects standard deviation. Image maps on the left side of the panels indicate p values from the t-tests for the respective time step, those on the right side indicate effect sizes (absolute Cohen's d_z). Black dotted lines on the left span time steps with significant differences between the mean trajectories.

tor items and trajectory divergence caused by two additional items sharing the reference color. This suggests that the increased attraction toward the relational pair seen here and in experiment six was not based on a generic interaction between multiple task-relevant items situated in close vicinity to each other. It thus strengthens the notion that the interaction observed in experiment six was indeed due to more complex processes of spatial language grounding in the case of an additional relational pair being present.

It has to be noted that while the double-item manipulation is more comparable to the presence of a full relational pair than are single items, it is still different with respect to the number of additional task-relevant colors. Since the only task-relevant colors are those of the target and the reference, no condition exists in which two additional task-relevant items are present that *differ* in color without forming a relational pair. In consequence, mechanisms associated with color representation cannot be excluded as the origin of the differential impact of a relational pair and double items. It is conceivable, for instance, that two same-colored items inhibit each other to a degree, leading to less activation and attraction overall, while activation and ensuing attraction caused by two differently colored items are not subject to this effect.

As concerns the remaining comparisons, the attraction toward the additional relational pair was stronger than other effects observed so far. It showed an onset comparable to the earliest onsets observed in previous experiments (but note that movement time tended to be high in this condition).

Attraction toward two distractor items as well was strong. Both the difference between mean trajectories and the effect size were larger than observed for single distractors in previous experiments (except compared to the effect size in experiment one). Thus, the attraction exerted by two close distractors seemed to combine into a larger effect, meaning that participants did not rule out these items as irrelevant at an early stage of processing.

The attraction toward two additional reference items was also present, as well suggesting that these items were not deemed irrelevant early on by the participants and played a role in the grounding process. However, compared to the effect of a single additional reference item in experiment six, which began particularly early, the effect here began at a relatively larger portion of total movement time (i.e., approximately 50%). Effect size was marginally larger compared to the reference effect in experiment six, while the difference between mean trajectories was similar. Thus, the reference effect did not appear to be as strongly affected by the double-item manipulation as the distractor effect. One might speculate that not all reference items in a display

were involved in grounding on every trial, thus showing a weaker average effect, which would be in line with the model, but it ultimately remains unclear why this pattern emerged.

In summary, experiment seven supports that the stronger attraction observed toward a relational pair was based on more complex grounding processes (with the caveat of a truly comparable double-item condition being unavailable). The pattern of attraction toward two identical items suggested that the effects of two same-colored items may be cumulative, although this seemed to be more strongly the case for the distractor than for the reference effect.

3.8 Discussion

In the mouse tracking paradigm described in this chapter, unknown spatial targets were specified by a relational phrase which had to be grounded in a visual scene to select the best-matching visual item with the computer mouse. The scenes were composed of a target item, one or two distractor items sharing the target color, a reference item, and zero, one, or two items sharing the reference color. Differently colored filler items were added at random positions such that each scene included twelve visual items in total.

Observed effects across experiments included the distractor effect, the reference effect, a spatial term effect, and increased attraction toward a competing relational pair.

The distractor effect was a trajectory bias to the side of the direct path on which one or two distractor items were located. In line with the spatial language model described in Chapter 2, the distractor effect is interpreted as arising from the attentional selection of items sharing the target feature, which occurs as a spatial phrase is grounded. That is, in the step of selecting a target item from among eligible candidates, a selection mechanism must at some point increase activation at the locations of these items through a form of feature attention to then select one of the activated items. The resulting increased activation within a map of visual space is assumed to affect the competition in maps where sensorimotor decisions are made, in line with the evidence described in the section *Embodiment in DFT (p.10)*.

The reference effect consisted of trajectory attraction toward the side of the direct path where the reference item of the spatial phrase, or items sharing its color, were located. In line with the model, the reference effect is interpreted as an impact of the attentional selection of items in reference color on motor

action. That is, in the step of selecting a reference item based on its color, a selection mechanism must at some point increase the saliency of the items sharing the reference color, to then select one.

The spatial term effect consisted of a bias with early onset into the direction described by the spatial term. Across experiments, it was shown to be evoked by both horizontal and vertical axis spatial terms. Its impact was likely sustained over large portions of the movements, although this could not be observed directly in most cases, presumably due to its overlap with the reference effect. In contrast to the distractor and the reference effect, the spatial term effect is interpreted along the lines of classical embodiment effects of language understanding, that is, as a biasing of motor movement by the semantic content of language (Tower-Richardi et al., 2012; Zwaan et al., 2012).

The increased attraction toward a competing relational pair was observed when two items in target and reference color were placed on one side of the direct path in vicinity to each other, such that they instantiated the opposite of the relation described in the spatial phrase. In experiment six, this took the form of the presence or absence of one item (e.g., the distractor) moderating the amount of additional attraction caused by adding the other item to the display (e.g., the additional reference); when one item was present, adding the other lead to a larger increase in attraction than when no other item was present. In experiment seven, the attraction caused by an additional relational pair was larger than the attraction exerted by two distractor items and larger than the attraction caused by two items in reference color. This is interpreted as showing that grounding processes performed in sensorimotor maps are flexibly organized and increase in complexity when more potential referents for a given spatial phrase are present.

This interpretation is in line with the model described in Chapter 2, in which additional grounding attempts occur when the incorrect items are selected in initial passes, with items being brought into the attentional foreground through feature attention in each grounding pass.

Effect onset and extent

Across experiments one, two, four, and five (which used the basic form of the paradigm with one distractor and one reference), the distractor effect tended to occur between approximately 30 and 40 percent of total movement time, and the reference effect between one and two thirds of total movement time. Relating this to absolute movement times yields a rough estimate of

between 300 and 600 milliseconds for the absolute time between display onset and effect onset (these numbers are deliberately kept vague and must be considered with care, as a possible covariation of absolute movement time and effect magnitude is not taken into account; for instance, if the effect estimate is dominated by trials with long movement times, then the absolute time between display and effect onset may be underestimated).

These effect onset times are broadly consistent with reaction times in visual search for color targets. For instance, Wolfe et al. (1990) found reaction times of approximately five to six hundred milliseconds for detecting a color target among up to 32 items in ten different colors (note that reaction times included the motor response); search was highly efficient with minimal reaction time slopes over increasing item number. Together, this suggests that items in the displays used in the experiments here could be distinguished very quickly via efficient visual search. It is thus likely that the observed effects were not affected by difficulties in finding the relevant items among fillers.

Relation to similar effects

The distractor and the reference effect are interpreted as arising from the evolution of attentional activation at different item locations during spatial language grounding. What alternative explanations could account for or contribute to the observed effects?

First, there is evidence that stimuli that capture attention attract movement trajectories (Moher et al., 2015; Wood et al., 2011). This is consistent with the interpretation of the distractor and reference effect as signatures of evolving attentional allocation.

Second, there is the rich body of mouse tracking research in which abstract cognitive tasks had to be solved and candidate solutions were linked to response locations in an arbitrary manner (e.g., Dale et al., 2007; Freeman et al., 2013; Barca & Pezzulo, 2012; Freeman & Ambady, 2011; see *Probing embodiment with mouse tracking*, p.16). Explaining the distractor effect along these lines would mean that which item satisfied the relational phrase best might have been computed in amodal substrates different from those in which sensorimotor decisions are made, and the candidate solutions, target and distractors, would then have been linked to item locations within a motor representation of visual space. In this view, the observed attraction would have arisen from evolution of task processing over time elsewhere.

A first argument against this interpretation is the fact that the visual displays in the present experiment appeared only at the time of movement on-

set and neither target nor distractor location was known in advance. The coupling of abstract task solutions to response locations would thus have occurred rapidly and in parallel to the response movement. Effect onsets occurred early after movement onset, especially in relation to the time required for visual search based on color targets, which leaves little time for such coupling to occur.

More importantly, the reference effect differs in nature from deviation toward candidate task solutions. The reference item was never a response option, so that the attraction toward it cannot be attributed to the same origin as that observed in previous mouse tracking studies. This, in turn, strengthens the notion that the distractor effect as well was based on processes in sensorimotor substrates rather than coupling abstract cognitive decisions to response locations.

Third, it has been shown that when a final reach target is marked only after movement onset, reach trajectories are biased in accordance with the distribution of precued potential target locations (Gallivan & Chapman, 2014; Chapman et al., 2010). This may have contributed to the distractor effect, considering that the final target was disambiguated only through grounding the relational phrase, which occurred after movement onset. As above, however, the reference effect does not fit into this picture, due to the reference item not being an action target as per the task instructions.

From the first three comparisons, it becomes clear that interpreting the effects observed here as signatures of cognitive processes that operate on sensorimotor substrates is strongly supported by the reference effect, while a distractor effect alone could more easily be interpreted along different lines. The finding of experiments six and seven, that a competing relational pair exerts a particularly strong effect, is another important source of support for the interpretation of the effects as based on grounding processes.

Fourth, there is evidence that reach trajectories can be attracted by color primes that share a prespecified target color and are presented briefly prior to the veridical target, but at positions incongruent with the final target location; for instance, a red prime flashed in an upper position gradually attracts a trajectory ultimately directed at a red target item in a lower position (Schmidt, 2002; Schmidt & Seydell, 2008). In other words, in those studies, participants momentarily moved toward a location that did not contain a target because the target-defining color was present at that location shortly before movement onset. Assuming that color words mentioned in a spatial phrase have a similar effect as task instructions in these studies, this may provide an expla-

nation for both distractor and reference attraction. An accessory assumption that must be made, however, is that two colors can be primed simultaneously, or it must be assumed that one took precedence in each trial. Moreover, the finding of greater attraction based on a competing relational pair makes further assumptions necessary to be explained in this manner. Although accessory assumptions would thus be required to explain the effects observed here through color priming, a contribution of such a mechanism cannot be ruled out.

Having said that, it is likely that the different experimental effects described in this section arose from a unified neural system that serves sensorimotor tasks and cognition. For instance, feature attention that serves to select eligible candidates, as in the model in Chapter 2, may naturally give rise to color priming. It lies in the nature of the neural mechanisms shared between different sensorimotor tasks that seemingly disparate effects observed in different experimental paradigms may not always be disentangled with respect to their origin. The similarity of the effects observed here to previously observed behavioral signatures can therefore be viewed as supporting rather than refuting the notion that spatial language grounding draws on modal substrates.

Relation to previous studies of spatial language

Relating the present findings to previous research in the area of understanding linguistic descriptions of spatial relations is possible only indirectly, because this research has typically focused on either eye tracking data, reaction time data from verification tasks, or acceptability ratings, rather than continuous behavioral measures. The most direct link lies in the fact that attraction to items potentially involved in a described relation is consistent with previous findings suggesting that attentional selection of items is required (Burigo & Knoeferle, 2015; Yuan et al., 2016; Franconeri et al., 2012).

Some previous studies have focused on the role of non-target and non-reference items (Logan & Compton, 1996; Carlson & Logan, 2001; Carlson & Hill, 2008). However, these studies did not specifically address the role of distractor items that are featurally identical to the target item. Instead, they focused on a different type of distracting items, namely ones that differed from both the target and the reference item with respect to feature conjunctions or complex object identity (e.g., using letters), so that efficient visual search was likely not possible (Wolfe & Horowitz, 2004). This means that, in these studies, determining the identity of items and thereby probing whether they

might be target or reference candidates may already have required attentional allocation.

The authors of those studies come to similar conclusions. Logan & Compton (1996), for instance, used letter displays for which a spatial relation between two letters had to be verified (e.g., "A above B?"). Adding irrelevant letters increased response times, which was attributed to visual search for items involved in the sought relation being difficult (Logan & Compton, 1996). This was confirmed by a similar study (Carlson & Logan, 2001) which also used a verification task and letter stimuli. This study additionally found that whether or not an added irrelevant letter matched the spatial term did not affect reaction times, concluding that the relation of the distractor to the reference item was not evaluated.

As noted above, these findings cannot easily be related to the present ones. In those previous studies, identifying irrelevant items as such probably required attentional allocation, but once they were identified it was immediately clear based on their identity that they were not the target or the reference, without the need to asses their relation to the reference. In other words, attentional focusing was forced, while relational assessment for the non-target was made unnecessary. In contrast, item displays used here likely allowed fillers to be easily distinguished from relevant items, while the distractor was featurally identical to the target and therefore needed to receive relational assessment in order to be ruled out as a target.

To the author's knowledge, no previous evidence exists that focuses on the effects of target-identical distractors in spatial language processing. It may be summarized that the current findings complement previous data in the area of spatial language understanding by demonstrating that distractors which are featurally identical to the target do receive an increased degree of attentional allocation, which may manifest in motor signatures.

Chapter 4

Conclusion

This thesis aimed to show that understanding language about the current environment is based on neural processes that operate on sensorimotor representations and are guided by linguistic input. This was examined for the case of grounding language about spatial relations in visual input, such as linking the phrase "The green object to the left of the red object" to its referents in an array of colored objects. This test case was chosen because spatial relations are abstract concepts in the sense that their ultimate referents — relations between objects — are not directly available in sensory input but can be constructed only through examining configurations of multiple objects.

A two-pronged approach was taken. First, a neural process model of the mechanisms that link spatial phrases to sensorimotor representations was developed. Second, a novel computer mouse tracking paradigm was devised and used to measure behavioral signatures of the hypothesized processes.

The neural dynamic model represents a prototype mechanism for how the linkage between modal sensorimotor and amodal linguistic substrates may be realized in a neural system. It is based on the theoretical framework of Dynamic Field Theory, in which metric patterns of activation that arise from sensorimotor surfaces are represented in continuous dynamic neural fields. Lateral neural interaction within these fields ensures a balance of stability and flexibility of the representational patterns, which is paramount for perception, motor action, and cognition in embodied, situated neural systems.

Model design was guided by the neural constraints associated with Dynamic Field Theory, by experimental evidence how humans ground spatial language, and by general constraints of attentional processing in biological nervous systems.

In the model, neural fields are complemented by dynamic neural nodes that represent the components of linguistic descriptions of spatial relations (spa-

tial phrases). Synaptic projections link the nodes to the fields, allowing them to affect the evolution of activation there, and thus control the grounding of spatial phrase components in visual input. The synaptic connectivity within the architecture realized computational steps that humans as well must perform when understanding spatial language.

The model is one seamless dynamical system, composed exclusively of neurally plausible building blocks. It is able to ground spatial language autonomously. By specifying the processes involved in spatial language grounding in a neurally plausible way, the model complements previous modeling approaches to spatial language grounding which have focused on isolated aspects of the required processes, or did not specify neural implementation details in a comparable manner.

It was shown how the model links each component in a spatial phrase to the corresponding perceptual objects in a visual scene. To this end, the evolution of activation in the neural substrates of the model was simulated for different combinations of spatial phrases and visual scenes through numerical solution of the differential equations that describe the model.

One important type of computational steps in these simulations involved selecting objects in the scene based on their visual features and bind them to the roles of target and reference in accord with the spatial phrase. Each of these steps required to increase neural activation in a representation of the visual scene at all positions where objects were located that shared the features denoted in the phrase. This was based on the neural mechanism of feature attention that was used to single out items of a given color.

The model was used as a heuristic to derive and interpret possible effects of spatial language grounding on behavioral measures. Specifically, in the experiments described in the thesis, the attentional selection of items was expected to become visible as a deviation of response trajectories toward those items. The expectation that processes in sensory substrates might influence motor action was based on a framework of embodied cognition that is closely associated with DFT and in agreement with previous neural and behavioral evidence.

Seven experiments assessed behavioral signatures of the grounding processes in sensorimotor substrates. The method of computer mouse tracking was applied in a novel paradigm, in which items as sources of behavioral effects could be located on either side of the path to the target item. Participants had to ground spatial phrases in visual scenes and select a described target item with the computer mouse.

The experiments measured trajectory deviations in the same space as that in which task processes operated. In this respect, they differed from previous mouse tracking experiments, in which solutions of abstract cognitive tasks, such as lexical decisions, were mapped to response locations in an arbitrary manner, so that task space and response space were separated.

Due to this novel aspect, the experiments described here allowed to observe the specification of response movements in dependency of the ongoing processes of spatial language grounding. Three major effects were found.

First, on their way to the ultimately selected target item, mouse trajectories deviated toward a distractor item that could be distinguished from the correct target only based on a lower goodness of fit with respect to the relational description. This was interpreted as a signature of the attentional selection of that item in the process of evaluating its match with the spatial term.

Second, an attraction to the reference item was observed, that is, toward the item relative to which the spatial terms, such as "left of", were defined. This was interpreted as showing that this item as well had to receive attentional allocation in the course of grounding a spatial phrase. Crucially, the reference item was never a behavioral target in the task. The effect thus disambiguates the origin of the behavioral signatures observed here from alternative explanations for the effects, such as movement direction being averaged over potential target positions, or candidate solutions competing in abstract processing and weighting potential targets in modal representations.

That processes associated with spatial language grounding were the origin of the observed attraction effects was furthermore supported by an experiment where in addition to the pair of items described in the phrase another pair of items was present that instantiated the opposite relation but shared the colors mentioned in the spatial phrase. It was shown that the competing relational pair attracted trajectories more strongly than what would have been expected based on the sum of effects observed when either item was presented alone. That the interaction was due to the items forming a relational pair was further supported by an experiment in which two items that were presented simultaneously but did not form a relational pair still exerted less attraction than two items that did form a relational pair.

The increased attraction toward a relational pair was interpreted as showing that items forming a relational pair lead to more complex grounding processes. That this could be observed in the trajectories lends further support to the notion that attraction effects in the experiments arose from flexible grounding processes in modal substrates rather than being generic effects of

attentional capture or epiphenomena of amodal processing.

Finally, across experiments, an effect of the spatial term was observed. It consisted of an early bias into the direction that the spatial term described and was independent of visual item positions. In contrast to the other findings, which showed process-based signatures, this effect is akin to classical effects of language embodiment insofar that motor action was biased in accordance with the semantic content of language.

Together, the experiments allowed to observe the online specification of response movements and its modulation by task processes of spatial language grounding. This modulation was leveraged here to support the claim that processes of language grounding do recruit sensorimotor systems rather than being performed on an abstract cognitive level. This interpretation is consistent with a stance on the embodiment of cognition that is linked to Dynamic Field Theory and holds that, like perception and action, cognition is tightly embedded in the sensory-motor loop.

It has initially been stated that much of the previous evidence in the domain of language embodiment has been correlational in nature, making it difficult to interpret in terms of supporting modal or amodal modes of language understanding. The present thesis has extended the available findings by showing not a general impact of language on perception or action in terms of facilitation or inhibition, but behavioral signatures of component processes implicated in spatial language grounding. Like previous evidence, this data cannot conclusively disprove that the core of language processing occurs on an abstract, amodal level, whose intermediate results are rapidly deployed to sensorimotor systems. However, the directed, process-based effects observed here require more specific assumptions to be explained in this manner than undirected facilitation or impedance effects.

In summary, this thesis has presented novel evidence for a grounded mode of language understanding, by focusing on a scenario in which the referents of the processed language are present in the sensory environment. The model complements this experimental evidence, by proposing how the required linkage between amodal linguistic information and continuous sensory information may be realized in a neural system. The relation of these findings to forms of non-situated grounded cognition about absent things and spaces will have to be clarified by future research. However, this thesis provided a step into the direction of a more concrete and process-oriented investigation of embodied higher cognition.

Bibliography

Abdi, H. (2007). The Bonferonni and Šidák corrections for multiple comparisons. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 103–107). Thousand Oaks, CA: Sage.

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

Anderson, S. E., Matlock, T., & Spivey, M. (2013). Grammatical aspect and temporal distance in motion descriptions. *Frontiers in Psychology*, *4*, 1–9.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, *34*(5), 502–508.

Barca, L., & Pezzulo, G. (2012). Unfolding visual lexical decision in time. *PLoS ONE*, *7*(4), e35932.

Barrós-Loscertales, A., González, J., Pulvermüller, F., Ventura-Campos, N., Bustamante, J. C., Costumero, V., ... Ávila, C. (2012). Reading salt activates gustatory brain regions: FMRI evidence for semantic grounding in a novel sensory modality. *Cerebral Cortex*, *22*(11), 2554–2563.

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

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617–45.

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, *2*(4), 716–724.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 245–284). Oxford: Oxford University Press.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84–91.

Bartolotti, J., & Marian, V. (2012). Language learning and control in monolinguals and bilinguals. *Cognitive Science*, *36*(6), 1129–1147.

Bastian, A., Riehle, A., Erlhagen, W., & Schöner, G. (1998). Prior information preshapes the population representation of movement direction in motor cortex. *NeuroReport*, *9*(2), 315–319.

Bastian, A., Schöner, G., & Riehle, A. (2003). Preshaping and continuous evolution of motor cortical representations during movement preparation. *The European Journal of Neuroscience*, *18*(7), 2047–2058.

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science*, *31*(5), 733–764.

Bicho, E., Mallet, P., & Schöner, G. (2000). Target representation on an autonomous vehicle with low-level sensors. *The International Journal of Robotics Research*, *19*(5), 424–447.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., & Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study. *Cognitive Brain Research*, *24*(3), 355–363.

Burigo, M., & Knoeferle, P. (2015). Visual attention during spatial language comprehension. *PLoS ONE*, *10*(1), e0115758.

Carlson, L. A., & Hill, P. L. (2008). Processing the presence, placement, and properties of a distractor in spatial language tasks. *Memory {&} Cognition*, *36*(2), 240–255.

Carlson, L. A., & Logan, G. D. (2001). Using spatial terms to select an object. *Memory and Cognition*, *29*(6), 883–892.

Carlson, L. A., & Logan, G. D. (2005). Attention and spatial language. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 330–336). San Diego, CA: Academic Press.

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

Carlson-Radvansky, L. A., & Irwin, D. E. (1993). Frames of reference in vision and language: Where is above? *Cognition*, *46*(3), 223–244.

Chambers, C., & Pressnitzer, D. (2014). Perceptual hysteresis in the judgment of auditory pitch shift. *Attention, Perception, and Psychophysics*, *76*(5), 1271–1279.

Chapman, C. S. (2011). *Using functional data analysis (fda) to analyze reach trajectories.* Retrieved 2017-09-01, from `http://www.per.ualberta.ca/acelab/wp-content/uploads/2014/09/Using-Functional-Data-Analysis_v1_april2011.pdf`

Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Culham, J. C., & Goodale, M. A. (2010). Reaching for the unknown: Multiple target encoding and real-time decision-making in a rapid reach task. *Cognition*, *116*(2), 168–176.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*(1), 1–22.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B*, *362*(1485), 1585–1599.

Cisek, P., & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. *Neuron*, *45*(5), 801–814.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, *33*(1), 269–298.

Cloutier, J., Freeman, J. B., & Ambady, N. (2014). Investigating the early stages of person perception: The asymmetry of social categorization by sex vs. age. *PLoS ONE*, *9*(1), e84677.

Coco, M. I., & Duran, N. D. (2016). When expectancies collide: Action dynamics reveal the interaction between stimulus plausibility and congruency. *Psychonomic Bulletin & Review*, *23*(6), 1920–1931.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Routledge.

Conway, B. R., & Tsao, D. Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. *Proceedings of the National Academy of Sciences*, *106*(42), 18034–18039.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology*, *107*(1), 84–107.

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*(1), 15–28.

Desai, R. H., Binder, J. R., Conant, L. L., & Seidenberg, M. S. (2010). Activation of sensory-motor areas in sentence comprehension. *Cerebral Cortex*, *20*(2), 468–478.

Dineva, E., & Schöner, G. (2018). How infants' reaches reveal principles of sensorimotor decision making. *Connection Science*, *30*(1), 53–80.

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, *110*(3), 412–431.

Dshemuchadse, M., Grage, T., & Scherbaum, S. (2015). Action dynamics reveal two types of cognitive flexibility in a homonym relatedness judgment task. *Frontiers in Psychology*, *6*, 1-10.

Dshemuchadse, M., Scherbaum, S., & Goschke, T. (2013). How decisions emerge: Action dynamics in intertemporal decision making. *Journal of Experimental Psychology: General*, *142*(1), 93–100.

Duran, N. D., Dale, R., & McNamara, D. S. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin and Review*, *17*(4), 486–491.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Erickson, R. (1974). Parallel "population" neural coding in feature extraction. In F. O. Schmitt & F. G. Worden (Eds.), *The neurosciences: Third study program* (pp. 155–169). Cambridge, MA: MIT Press.

Erlhagen, W., Bastian, A., Jancke, D., Riehle, A., & Schöner, G. (1999). The distribution of neuronal population activation (DPA) as a tool to study interaction and integration in cortical representations. *Journal of Neuroscience Methods*, *94*(1), 53–66.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572.

Fahim, M., & Rezanejad, A. (2014). An introduction to embodied cognition. *International Journal of Language and Linguistics*, *2*(4), 283–289.

Farmer, T. A., Anderson, S. E., & Spivey, M. J. (2007). Gradiency and visual context in syntactic garden-paths. *Journal of Memory and Language*, *57*(4), 570–595.

Farmer, T. A., Cargill, S. A., Hindy, N. C., Dale, R., & Spivey, M. J. (2007). Tracking the continuity of language comprehension: Computer mouse trajectories suggest parallel syntactic processing. *Cognitive Science*, *31*(5), 889–909.

Farmer, T. A., Liu, R., Mehta, N. S., & Zevin, J. D. (2009). Native language experience influences the perceived similarity of second language vowel categories. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 2588–2593).

Faubel, C., & Zibner, S. K. (2010). A neuro-dynamic object recognition architecture enhanced by foveal vision and a gaze control mechanism. *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems (IROS)*, 1171–1176.

Flumini, A., Barca, L., Borghi, A. M., & Pezzulo, G. (2014). How do you hold your mouse? Tracking the compatibility effect between hand posture and stimulus size. *Psychological Research*, *79*(6), 928–938.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior research methods*, *42*(1), 226–41.

Freeman, J. B., & Ambady, N. (2011). Hand movements reveal the time-course of shape and pigmentation processing in face categorization. *Psychonomic Bulletin and Review*, *18*(4), 705–712.

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology: General*, *137*(4), 673–690.

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods*, *45*(1), 83–97.

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 1–6.

Freeman, J. B., Ma, Y., Han, S., & Ambady, N. (2013). Influences of culture and visual context on real-time social categorization. *Journal of Experimental Social Psychology*, *49*(2), 206–210.

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652–654.

Gallivan, J. P., & Chapman, C. S. (2014). Three-dimensional reach trajectories as a probe of real-time decision-making between multiple competing targets. *Frontiers in Neuroscience*, *8*, 1–19.

Gasser, M. (2004). The origins of arbitrariness in language. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the cognitive science society* (Vol. 26, p. 434-439). Mahwah, NJ: Erlbaum.

Gentilucci, M., Benuzzi, F., Bertolani, L., Daprati, E., & Gangitano, M. (2000). Language and motor control. *Experimental Brain Research*, *133*(4), 468–490.

Georgopoulos, A. P., Caminiti, R., Kalaska, J. F., & Massey, J. T. (1983). Spatial coding of movement: A hypothesis concerning the coding of movement direction by motor cortical populations. *Experimental Brain Research*, *49*(7), 327–336.

Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., & Massey, J. T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience*, *2*(11), 1527–1537.

Georgopoulos, A. P., Kettner, R. E., & Schwartz, A. B. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *Journal of Neuroscience*, *8*(8), 2928–2937.

Georgopoulos, A. P., Schwartz, A., & Kettner, R. (1986). Neuronal population coding of movement direction. *Science*, *233*(4771), 1416–1419.

Georgopoulos, A. P., Taira, M., & Lukashin, A. (1993). Cognitive neurophysiology of the motor cortex. *Science*, *260*(5104), 47–52.

Ghez, C., Favilla, M., Ghilardi, M. F., Gordon, J., Bermejo, R., & Pullman, S. (1997). Discrete and continuous planning of hand movements and isometric force trajectories. *Experimental Brain Research*, *115*(2), 217–233.

Glenberg, A. M. (1997). What memory is for. *The Behavioral and Brain Sciences*, *20*(1), 1–19.

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

Glover, S., & Dixon, P. (2002). Semantics affect the planning but not control of grasping. *Experimental Brain Research*, *146*(3), 383–387.

Goldinger, S. D., Papesh, M. H., Barnhart, A. S., Hansen, W. A., & Hout, M. C. (2016). The poverty of embodied cognition. *Psychonomic Bulletin and Review*, *23*(4), 959–978.

Gotts, S. J., Milleville, S. C., Bellgowan, P. S., & Martin, A. (2011). Broad and narrow conceptual tuning in the human frontal lobes. *Cerebral Cortex*, *21*(2), 477–491.

Groh, J. M., Born, R. T., & Newsome, W. T. (1997). How is a sensory map read out? Effects of microstimulation in visual area MT on saccades and smooth pursuit eye movements. *The Journal of Neuroscience*, *17*(11), 4312–4330.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In F. M. Snell & R. Rosen (Eds.), *Progress in theoretical biology, vol. 5* (pp. 233–374). New York: Academic press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70–84.

Hartigan, P. M. (1985). Algorithm AS 217: Computation of the dip statistic to test for unimodality. *Applied Statistics*, *34*(3), 320.

Hayward, W. G., & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, *55*(1), 39–84.

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, *18*(3), 384–401.

Hock, H. S., Kogan, K., & Espinoza, J. K. (1997). Dynamic, state-dependent thresholds for the perception of single-element apparent motion: bistability from local cooperativity. *Perception & Psychophysics*, *59*(7), 1077–1088.

Hock, H. S., & Schöner, G. (2010). A neural basis for perceptual dynamics. In R. Huys & V. K. Jirsa (Eds.), *Nonlinear dynamics in human behavior* (pp. 151–177). Berlin and Heidelberg: Springer.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

Hyun, J.-S., Woodman, G. F., & Luck, S. J. (2009). The role of attention in the binding of surface features to locations. *Visual Cognition*, *17*(1-2), 10–24.

Jancke, D., Erlhagen, W., Dinse, H. R., Akhavan, A. C., Giese, M., Steinhage, A., & Schöner, G. (1999). Parametric population representation of retinal location: Neuronal interaction dynamics in cat primary visual cortex. *The Journal of Neuroscience*, *19*(20), 9016–9028.

Johnson, J. S., & Simmering, V. R. (2015). Integrating perception and working memory in a three-layer dynamic field architecture. In G. Schöner, J. Spencer, & the DFT Research Group (Eds.), *Dynamic thinking: A primer on dynamic field theory* (pp. 151–168). New York, NY: Oxford University Press.

Johnson, J. S., Spencer, J. P., Luck, S. J., & Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*(5), 568–577.

Johnson, J. S., Spencer, J. P., & Schöner, G. (2009). A layered neural architecture for the consolidation, maintenance, and updating of representations in visual working memory. *Brain Research*, *1299*, 17–32.

Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1187–1211.

Kaschak, M. P., & Borreggine, K. L. (2008). Temporal dynamics of the action-sentence compatibility effect. *Quarterly Journal of Experimental Psychology*, *61*(6), 883–895.

Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R. H., Aveyard, M., Blanchard, A. A., & Zwaan, R. A. (2005). Perception of motion affects language processing. *Cognition*, *94*(3), 79–89.

Kiefer, M., Sim, E.-J., Herrnberger, B., Grothe, J., & Hoenig, K. (2008). The sound of concepts: Four markers for a link between auditory and conceptual brain systems. *Journal of Neuroscience*, *28*(47), 12224–12230.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1–16.

Kluth, T., Burigo, M., & Knoeferle, P. (2016). Shifts of attention during spatial language comprehension: A computational investigation. In *International Conference on Agents and Artificial Intelligence* (Vol. 2, pp. 213–222). SCITEPRESS – Science and Technology Publications, Lda.

Koop, G. J., & Johnson, J. G. (2011). Response dynamics: A new window on the decision process. *Judgment and Decision Making*, *6*(8), 750–758.

Krakauer, J. W., Pine, Z. M., Ghilardi, M. F., & Ghez, C. (2000). Learning of visuomotor transformations for vectorial planning of reaching trajectories. *Journal of Neuroscience*, *20*(23), 8916–8924.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York, NY: Basic Books.

Lee, C., Rohrer, W. H., & Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, *332*(6162), 357–360.

Lepora, N. F., & Pezzulo, G. (2015). Embodied choice: How action influences perceptual decision making. *PLoS Computational Biology*, *11*(4), 1–22.

Lins, J., & Schöner, G. (2014). A neural approach to cognition based on dynamic field theory. In S. Combes, P. beim Graben, R. Potthast, & J. Wright (Eds.), *Neural fields* (pp. 319–339). Berlin, Heidelberg: Springer.

Lins, J., & Schöner, G. (2017). Mouse tracking shows attraction to alternative targets while grounding spatial relations. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2586–2591). Austin, TX: Cognitive Science Society.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neuro-behavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*(6), 1490–1511.

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

Logan, G. D., & Compton, B. J. (1996). Distance and distraction effects in the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 159–172.

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

Lomp, O., Richter, M., Zibner, S. K. U., & Schöner, G. (2016). Developing dynamic field theory architectures for embodied cognitive systems with cedar. *Frontiers in Neurorobotics*, *10*, 1–18.

Magnuson, J. S. (2005). Moving hand reveals dynamics of thought. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 9995–9996.

Mahon, B. Z. (2015). What is embodied about cognition? *Language, Cognition and Neuroscience*, *30*(4), 420–429.

Mechler, F. (2002). *Hartigan's dip statistic.* Retrieved 2017-07-05, from `http://nicprice.net/diptest/`

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788–804.

Meyer, K., & Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends in Neurosciences*, *32*(7), 376–382.

Moher, J., Sit, J., & Song, J.-H. (2015). Goal-directed action is automatically biased towards looming motion. *Vision Research*, *113*, 188-197.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 1-12.

Moore, C. M., Elsinger, C. L., & Lleras, A. (2001). Visual attention and the apprehension of spatial relations: The case of depth. *Perception & Psychophysics*, *63*(4), 595–606.

Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, *37*(1), 149–157.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Ottes, F. P., van Gisbergen, J. A., & Eggermont, J. J. (1985). Latency dependence of colour-based target vs nontarget discrimination by the saccadic system. *Vision Research*, *25*(6), 849–862.

Pasupathy, A., & Connor, C. E. (2001). Shape representation in area V4: Position-specific tuning for boundary conformation. *Journal of Neurophysiology*, *86*(5), 2505–2519.

Pecher, D., Boot, I., & Dantzig, S. V. (2011). Abstract concepts: Sensory-motor grounding, metaphors, and beyond. *Psychology of Learning and Motivation*, *54*, 217–248.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Pfister, R., Schwarz, K. A., Janczyk, M., Dale, R., & Freeman, J. B. (2013). Good things peak in pairs: A note on the bimodality coefficient. *Frontiers in Psychology*, *4*, 1-4.

Pulvermüller, F., Shtyrov, Y., & Ilmoniemi, R. (2005). Brain signatures of meaning access in action word recognition. *Journal of Cognitive Neuroscience*, *17*(6), 884–892.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, *3*(1), 111–132.

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

Richardson, D. C., Spivey, M. J., Barsalou, L. W., & Mcrae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science*, *27*, 767–780.

Richardson, D. C., Spivey, M. J., Edelman, S., & Naples, A. J. (2001). "Language is spatial": Experimental evidence for image schemas of concrete and abstract verbs. In *23rd Annual Conference of the Cognitive Science Society* (pp. 845–860). Mawhah, NJ: Erlbaum.

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In P. Bello, M. Guarini, M.McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2847–2852). Austin, TX: Cognitive Science Society.

Richter, M., Lins, J., Schneegans, S., & Schöner, G. (2014). A neural dynamic architecture resolves phrases about spatial relations in visual scenes. In S. Wermter & et al. (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2014. Lecture Notes in Computer Science* (Vol. 8681, pp. 201–208). Springer.

Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Sciences*, *9*(1), 35–47.

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition.

In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 2457–2464).

Richter, T., & Zwaan, R. A. (2009). Processing of color words activates color representations. *Cognition*, *111*(3), 383–389.

Salverda, A. P., & Tanenhaus, M. K. (2017). The visual world paradigm. In A. M. B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics: A practical guide* (pp. 89–110). Malden, MA: Wiley-Blackwell.

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

SAS Institute. (2012). *Sas/stat 12.1 user's guide: Survival analysis*. Cary, NC: SAS Institute Inc.

Scherbaum, S., Dshemuchadse, M., Leiberg, S., & Goschke, T. (2013). Harder than expected: Increased conflict in clearly disadvantageous delayed choices in a computer game. *PLoS ONE*, *8*(11), e79310.

Scherbaum, S., Frisch, S., Leiberg, S., Lade, S. J., Goschke, T., & Dshemuchadse, M. (2016). Process dynamics in delay discounting decisions: An attractor dynamics approach. *Judgement and Decision Making*, *11*(5), 472–495.

Scherbaum, S., Gottschalk, C., Dshemuchadse, M., & Fischer, R. (2015). Action dynamics in multitasking: The impact of additional task factors on the execution of the prioritized motor movement. *Frontiers in Psychology*, *6*, 1–8.

Scherbaum, S., & Kieslich, P. J. (2017). Stuck at the starting line: How the starting procedure influences mouse-tracking data. *Behavior Research Methods*, 1–14.

Schmidt, T. (2002). The finger in flight: Real-time motor control by visually masked color stimuli. *Psychological Science*, *13*(2), 112–118.

Schmidt, T., & Seydell, A. (2008, apr). Visual attention amplifies response priming of pointing movements to color targets. *Perception {&} Psychophysics*, *70*(3), 443–455.

Schneegans, S. (2016). *Dynamic field theory of visuospatial cognition* (doctoral dissertation). Ruhr-Universität Bochum, Universitätsbibliothek.

Schneegans, S., Lins, J., & Schöner, G. (2015). Embedding Dynamic Field Theory in Neurophysiology. In G. Schöner, J. Spencer, & the DFT Research Group (Eds.), *Dynamic thinking: A primer on dynamic field theory* (pp. 61–93). New York, NY: Oxford University Press.

Schneegans, S., Lins, J., & Spencer, J. P. (2015). Integration and selection in multidimensional dynamic fields. In G. Schöner, J. Spencer, & the DFT Research Group (Eds.), *Dynamic thinking: A primer on dynamic field theory* (pp. 121–149). New York, NY: Oxford University Press.

Schneegans, S., & Schöner, G. (2008). Dynamic field theory as a framework for understanding embodied cognition. In P. Calvo & T. Gomila (Eds.), *Handbook of cognitive science: An embodied approach* (pp. 241–271). San Diego, CA: Elsevier.

Schneegans, S., & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109.

Schneegans, S., Spencer, J. P., & Schöner, G. (2015). Integrating "what" and "where": Visual working memory for objects in a scene. In G. Schöner, J. Spencer, & the DFT Research Group (Eds.), *Dynamic thinking: A primer on dynamic field theory* (pp. 197–226). New York, NY: Oxford University Press.

Schöner, G. (2008). Dynamical systems approaches to cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 101–126). Cambridge, UK: Cambridge University Press.

Schöner, G., Faubel, C., Dineva, E., & Bicho, E. (2015). Embodied neural dynamics. In G. Schöner, J. Spencer, & the DFT Research Group (Eds.), *Dynamic thinking: A primer on dynamic field theory* (pp. 95–118). New York: Oxford University Press.

Schöner, G., Reimann, H., & Lins, J. (2015). Neural dynamics. In G. Schöner, J. Spencer, & the DFT Research Group (Eds.), *Dynamic thinking: A primer on dynamic field theory* (pp. 5–34). New York, NY: Oxford University Press.

Schöner, G., Spencer, J. P., & the DFT Research Group. (2015). *Dynamic thinking: A primer on dynamic field theory*. New York, NY: Oxford University Press.

Schultheis, H. (2007). A computational model of control mechanisms in spatial term use. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the*

29th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Schultheis, H., & Barkowsky, T. (2011). Casimir: An architecture for mental spatial knowledge processing. *Topics in Cognitive Science*, *3*(4), 778–795.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Sell, A. J., & Kaschak, M. P. (2011). Processing time shifts affects the execution of motor responses. *Brain and Language*, *117*(1), 39–44.

Sherrington, C. S. (1906). *The integrative action of the nervous system*. New Haven, CT: Yale University Press.

Song, J.-H., & Nakayama, K. (2006). Role of focal attention on latencies and trajectories of visually guided manual pointing. *Journal of Vision*, *6*(9), 982–995.

Song, J.-H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences.*, *13*(8), 360–366.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*(2), 153–156.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tekülve, J., Zibner, S. K. U., & Schöner, G. (2016). A neural process model of learning to sequentially organize and activate pre-reaches. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 318–325).

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., ... Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, *17*(2), 273–281.

Tower-Richardi, S., Brunye, T., Gagnon, S., Mahoney, C., & Taylor, H. (2012). Abstract spatial concept priming dynamically influences real-world actions. *Frontiers in Psychology*, *3*, 1–12.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Ts'o, D., Gilbert, C., & Wiesel, T. (1986). Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. *Journal of Neuroscience*, *6*(4), 1160–1170.

Vavrečka, M., & Farkaš, I. (2014). A multimodal connectionist architecture for unsupervised grounding of spatial language. *Cognitive Computation*, *6*(1), 101–112.

Wifall, T., Buss, A. T., Farmer, T. A., Spencer, J. P., & Hazeltine, E. (2017). Reaching into response selection: Stimulus and response similarity influence central operations. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 555–568.

Wilimzig, C., Schneider, S., & Schöner, G. (2006). The time course of saccadic decision making: Dynamic field theory. *Neural Networks*, *19*(8), 1059–1074.

Willems, R. M., & Francken, J. C. (2012). Embodied cognition: Taking the next step. *Frontiers in Psychology*, *3*, 1–3.

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, *12*(1), 1–24.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York, NY: McGraw-Hill.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501.

Wolfe, J. M., Yu, K. P., Stewart, M. I., Shorter, A. D., Friedman-Hill, S. R., & Cave, K. R. (1990). Limitations on the parallel guidance of visual search: color × color and orientation × orientation conjuctions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 879–892.

Wood, D. K., Gallivan, J. P., Chapman, C. S., Milne, J. L., Culham, J. C., & Goodale, M. A. (2011). Visual salience dominates early visuomotor competition in reaching behavior. *Journal of Vision*, *11*(10), 1–11.

Yaxley, R. H., & Zwaan, R. A. (2007). Simulating visibility during language comprehension. *Cognition*, *105*, 229–236.

Yuan, L., Uttal, D., & Franconeri, S. (2016). Are categorical spatial relations encoded by shifting visual attention between objects? *PLoS ONE*, *11*(10), e0163141.

Zibner, S. K. U. (2017). *A neuro-dynamic architecture for autonomous visual scene representation* (doctoral dissertation). Ruhr University Bochum, Verlag Dr. Hut.

Zibner, S. K. U., Faubel, C., & Schöner, G. (2011). Making a robotic scene representation accessible to feature and label queries..

Zibner, S. K. U., Tekülve, J., & Schöner, G. (2015). The neural dynamics of goal-directed arm movements: A developmental perspective. *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics, (ICDL-EpiRob)*, 154–161.

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*(5), 229–234.

Zwaan, R. A., Madden, C. J., Yaxley, R. H., & Aveyard, M. E. (2004). Moving words: Dynamic representations in language comprehension. *Cognitive Science*, *28*(4), 611–619.

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PLoS ONE*, *7*(12), e51382.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*(2), 168–171.

Zwaan, R. A., van der Stoep, N., Guadalupe, T., & Bouwmeester, S. (2012). Language comprehension in the balance: The robustness of the action-compatibility effect (ACE). *PLoS ONE*, *7*(2), e31204.

Appendices

A Participant data questionnaire

Probandenfragebogen

Probandencode: _____

Alter: _____

Geschlecht: O Weiblich O Männlich

Leiden Sie unter einer Form von Farbenblindheit?

O Nein

O Rot-Grün-Blindheit oder -Schwäche

O Blau-Blindheit oder –Schwäche

O Vollständige Farbenblindheit

Leiden Sie unter einer anderen Sehschwäche (z.B. Kurzsichtigkeit)?

O Ja O Nein

Falls JA, ist diese vollständig durch eine Sehhilfe korrigiert (z.B. Brille)?

O Ja O Nein

Ist Deutsch Ihre Muttersprache?

O Ja O Nein

Falls NEIN, in welchem Alter haben Sie Deutsch gelernt?

_____

B Informed consent form

INSTITUT FÜR NEUROINFORMATIK – RUHR-UNIVERSITÄT BOCHUM – UNIVERSITÄTSSTR. 150 – 44801 BOCHUM

Information & Einverständniserklärung

Für Teilnehmer an der Studie
„Manuelle Auswahl von verbal beschriebenen Zielobjekten".

Ziel der Studie. Die Studie untersucht zielorientierte Bewegungen auf sprachlich beschriebene Objekte. Sie soll Hinweise liefern, wie das Gehirn visuelle und sprachliche Informationen verbindet.

Ablauf. Zunächst werden einige allgemeine Daten per Fragebogen erhoben. Während des Experiments sitzen Sie vor einem Bildschirm, auf dem eine Zielbeschreibung und anschließend mehrere Objekte gezeigt werden. Ihre Aufgabe ist es, mit der Maus das beschriebene Objekt auszuwählen. Das Experiment umfasst mehrere solcher Durchgänge.

Dauer und Vergütung. Die Gesamtdauer der Studie beträgt etwa 60 Minuten. Sie erhalten eine Aufwandsentschädigung von € 10 in bar.

Risiken. Die Teilnahme an der Studie ist nicht mit besonderen Risiken verbunden.

Vertraulichkeit. Die in Fragebogen und Experiment erhobenen Daten werden anonymisiert, so dass sie nicht mit Ihren persönlichen Daten in Verbindung gebracht werden können.

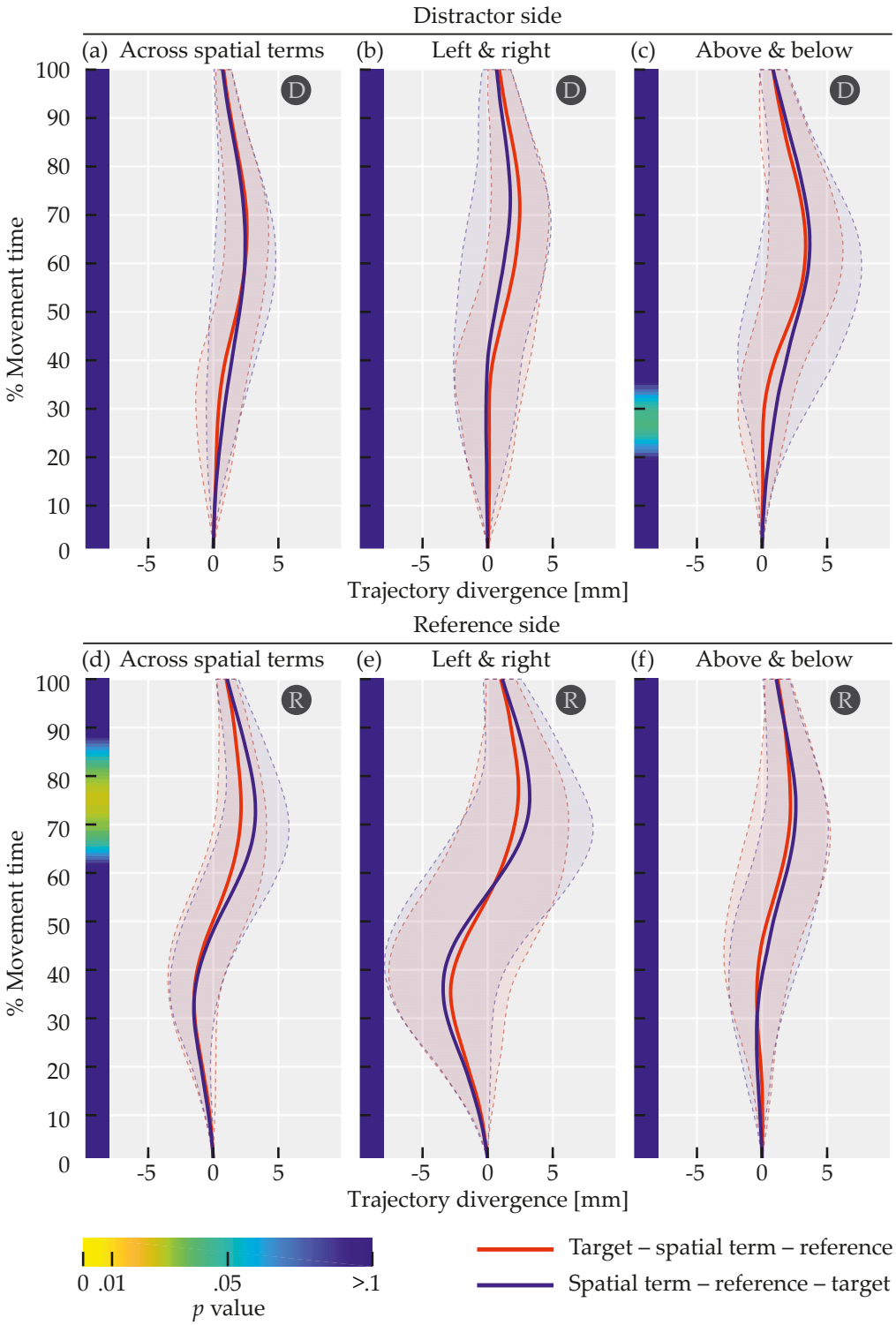
Freiwilligkeit. Die Teilnahme ist freiwillig. Falls Sie sich für die Teilnahme entscheiden, bitten wir Sie, diese Einverständniserklärung zu unterschreiben. Auch danach können Sie zu jeder Zeit Ihr Einverständnis widerrufen und das Experiment ohne Angabe von Gründen abbrechen.

Einverständniserklärung

Ich habe die voranstehenden Informationen gelesen und hatte Gelegenheit verbleibende Fragen mit dem Versuchsleiter zu klären. Ich erkläre mich freiwillig bereit an der o.g. Studie teilzunehmen.

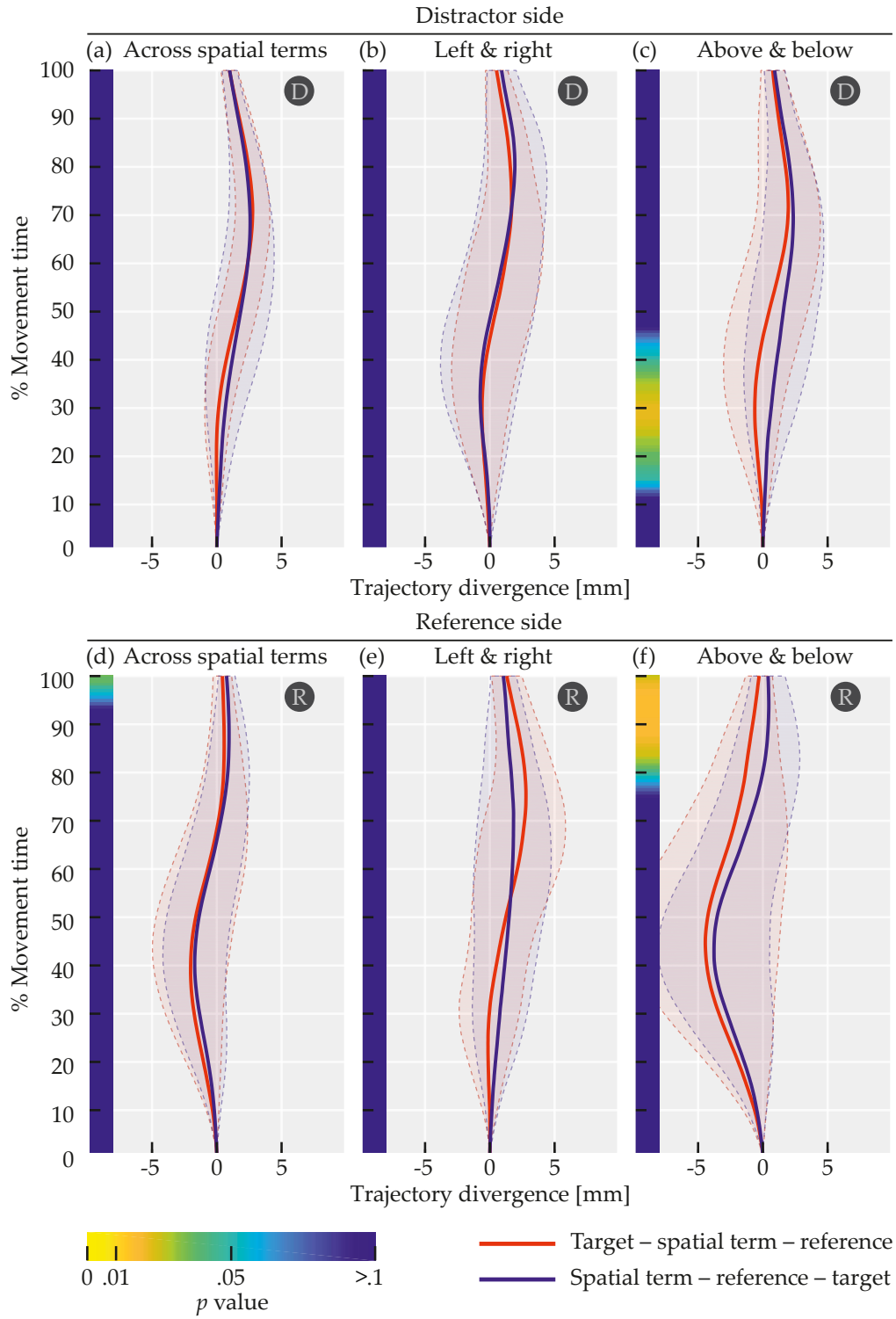
Teilnehmer:	Versuchsleiter:
_____	_____
Vor- und Nachname	Vor- und Nachname
_____	_____
Datum, Unterschrift	Datum, Unterschrift

C Comparisons by word order for experiment four



Comparisons of mean divergence (difference scores between item of interest right minus left) between word orders for experiment four. Gray circles labeled 'D' or 'R' indicate the effective side of the distractor or reference; curve biases toward the side of these circles indicate divergence consistent with a bias into the respective item's direction. Transparent regions delimited by dashed lines indicate standard deviation between participant means. Image maps indicate p values from t-tests at that time step. There were no significant time steps.

D Comparisons by word order for experiment five



Comparisons of mean divergence (difference scores between item of interest right minus left) between word orders for experiment five. Gray circles labeled 'D' or 'R' indicate the effective side of the distractor or reference; curve biases toward the side of these circles indicate divergence consistent with a bias into the respective item's direction. Transparent regions delimited by dashed lines indicate standard deviation between participant means. Image maps indicate p values from t-tests at that time step. There were no significant time steps.