Does Direction Matter? Linguistic Asymmetries Reflected in Visual Attention[☆]

Thomas Kluth[a,*], Michele Burigo[a], Holger Schultheis[b], Pia Knoeferle[c]

[a] *CITEC (Cognitive Interaction Technology Excellence Cluster), Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany*
[b] *Bremen Spatial Cognition Center, University of Bremen, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany*
[c] *Berlin School of Mind and Brain, Einstein Center for Neuroscience Berlin, and*
Department of German Studies and Linguistics, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany

Abstract

Language and vision interact in non-trivial ways. Linguistically, spatial utterances are often asymmetrical as they relate more stable objects (reference objects) to less stable objects (located objects). Researchers have claimed that such linguistic asymmetry should also be reflected in the allocation of visual attention when people process a depicted spatial relation described by spatial language. More specifically, it was assumed that people move their attention from the reference object to the located object. However, recent theoretical and empirical findings challenge the directionality of this attentional shift. In this article, we present the results of an empirical study based on predictions generated by computational cognitive models implementing different directionalities of attention. Moreover, we thoroughly analyze the computational models. While our results do not favor any of the implemented directionalities of attention, we found that two unknown sources of geometric information affect spatial language understanding. We provide modifications to the computational models that substantially improve their performance on empirical data.

Keywords: language and vision · spatial language · spatial relations · visual attention · cognitive modeling

1. Introduction

Speaking about things in our environment requires the integration of many different processes and representations (perceptual and cognitive) in a matter of seconds. During such interaction, visual processes affect linguistic processes and linguistic processes affect visual processes (Anderson, Chiu, Huette, & Spivey, 2011). Spatial language processing offers a flourishing field to investigate this reciprocal interaction (e.g., Carlson-Radvansky & Irwin, 1993; Crawford, Regier, & Huttenlocher, 2000; Hayward & Tarr, 1995; Landau & Jackendoff, 1993). In particular, focused visual attention appears necessary for evaluating linguistic descriptions of given visual spatial relations (e.g., Carlson & Logan, 2005; Logan, 1995; Logan & Sadler, 1996). In this article, we focus on the role of visual attention but also object distance and geometric object properties for the processing of spatial language.

1.1. Spatial Language Processing

Consider a scene with a spatial relation between two objects, say, a bike and a house (cf. Talmy, 2000, p. 183) and a related sentence (1).

(1) The bike is in front of the house.
(2) The house is behind the bike.

Cognitive linguists have claimed that spatial utterances such as in (1) are asymmetric (e.g., Landau & Jackendoff, 1993), i.e., the roles of the bike and the house are not easily interchangeable: Saying (2) is formally correct but rarely heard in everyday communication. The roles of the linguistic entities in spatial utterances are the 'reference object' (or ground, landmark, reference, relatum; the house in example (1)) versus the 'located object' (or figure, trajector, target, locatum; the bike in example (1)). The reference object (RO) is assumed to be less movable and larger than the located object (LO; e.g., Landau & Jackendoff, 1993; Talmy, 2000) and most spatial language researchers have focused on investigating the properties of the RO (e.g., Carlson-Radvansky & Logan, 1997; O'Keefe, 2003; Regier & Carlson, 2001; but see Burigo & Sacchi, 2013). However, a spatial utterance should help the hearer to find the LO such that she can attend to it. This motivated the claim that "the viewer's attention should move *from* the reference object *to* the located object" (p. 499, Logan & Sadler, 1996, emphasis in the original; see also Logan, 1995, p. 115, and Logan & Zbrodoff, 1999, p. 72).

[☆] Preliminary results of this work have been presented at the 5. interdisziplinärer Workshop Kognitive Systeme (Kluth, Burigo, & Knoeferle, 2016a; PSP analysis), the 6th UK Cognitive Linguistics Conference (Kluth, Burigo, & Knoeferle, 2016b; empirical data), the 14th International Conference on Cognitive Modeling (Kluth, Burigo, Schultheis, & Knoeferle, 2016a; PSP, empirical data, landscaping, and MFA), the 13th Conference of the German Society for Cognitive Science (Kluth, Burigo, Schultheis, & Knoeferle, 2016b; empirical data, AVS-BB and rAVS-CoO model), and the 10th Embodied and Situated Language Processing Conference (Kluth, Burigo, Schultheis, & Knoeferle, 2017, empirical data). This article provides additional and more in-depth analyses of the results.

[*] Corresponding author
Email addresses: tkluth@cit-ec.uni-bielefeld.de (Thomas Kluth), mburigo@cit-ec.uni-bielefeld.de (Michele Burigo), schulth@uni-bremen.de (Holger Schultheis), pia.knoeferle@hu-berlin.de (Pia Knoeferle)

Since focused visual attention appears necessary for relating spatial descriptions to depicted spatial relations (see Carlson & Logan, 2005, for a review), this claim has influenced the research on spatial language. One example of this tacit acknowledgment of the theorized directionality of attention is the Attentional Vector Sum (AVS) model by Regier and Carlson (2001). The input for the AVS model is a 2-dimensional spatial configuration of a RO and a LO (e.g., a point above a rectangle) as well as a spatial preposition (e.g., *above*). The output of the AVS model is an acceptability score, i.e., how well the spatial preposition describes the spatial configuration.

The prime motivation of Regier and Carlson (2001) was to investigate "[...] what perceptual or cognitive structures are reflected in these linguistic judgments [spatial language evaluation]? Does spatial perception shape spatial language in this instance, and if so, how?" (p. 274). Regier and Carlson (2001) successfully identified a mechanism that accounted considerably better for human spatial language acceptability scores than alternative mechanisms: an attentional vector sum. Regier and Carlson (2001) motivated the attentional vector sum with two observations. First, processing of spatial relations requires focal attention (i.e., spatial relations do not "pop-out" in a visual search task; Logan, 1994, 1995). Second, the representation of direction (i) in monkeys' motor cortex during arm movements (Georgopoulos, Schwartz, & Kettner, 1986), (ii) for saccadic eye movements (Lee, Rohrer, & Sparks, 1988), and (iii) in motion perception (Wilson & Kim, 1994) is best described by a weighted vector sum across populations of neurons. Indeed, the combination of a vector sum weighted with an attentional distribution in the AVS model outperformed several competing models. Although Regier and Carlson (2001) did not focus on the *directionality* of attention, their AVS model implicitly incorporates a directionality of attention from the RO to the LO via the direction of the vectors in the model (i.e., a movement of attention from the house to the bike in "The bike is in front of the house.").

Lipinski, Schneegans, Sandamirskaya, Spencer, and Schöner (2012) presented a comparatively more fine-grained model that is compatible with both neuronal mechanisms and the mechanisms assumed by the AVS model (see Richter, Lins, Schneegans, Sandamirskaya, & Schöner, 2014; Richter, Lins, & Schöner, 2016, 2017, for extensions to the model). In their model, the representation of the RO is activated prior to the representation of the LO suggesting that the directionality of attention goes from the RO to the LO (see in particular Richter et al., 2016, 2017).[1]

In related empirical work, the AVS model is also interpreted as implementing an "[...] attention allocation from a RO to a LO" (Coventry et al., 2010, p. 211). However, the same study found that for superior prepositions (*over/above*) people mostly fixated first the LO and next the RO, indicating a reversed directionality of overt attention. Since Coventry et al. (2010) recorded eye movements after listeners heard the spatial utterance, we cannot directly time-lock these attentional patterns to the unfolding interpretation or to the processing of individual words. Nevertheless, the patterns are compatible with findings that people inspect objects in the order they are mentioned. When the LO is mentioned before the RO, people first fixate the LO more than the RO followed by more looks to the RO compared with the LO (Burigo & Knoeferle, 2015).

Computational models developed to enable robots to interpret spatial language also follow this order (first attending to the LO then attending to the RO, Roy & Mukherjee, 2005). This fits with early results from Huttenlocher and Strauss (1968) and a related study summarized in Landau and Jackendoff (1993, p. 225): Both children and adults responded faster to instructions when they mentioned LO-like objects first (e.g., movable blocks) than RO-like objects (e.g., non-movable blocks). Perhaps sentences are more readily turned into action if the LO is mentioned first.

Recently, Franconeri, Scimeca, Roth, Helseth, and Kahn (2012) proposed that people must shift their attention in order to encode spatial relations and that the *direction* of that shift matters for the processing of spatial relations. Based on this "shift account", Roth and Franconeri (2012) found that participants were quicker to judge a spatial language question when the LO appeared slightly before the RO on the screen, forcing them to shift their attention from the LO to the RO. Interestingly, the direction of this shift is not intuitive: After hearing "circle above rectangle", participants' attention shifted from the top object (the LO) to the bottom object (the RO), i.e., in the opposite direction of the spatial preposition ("This flip is counterintuitive, but certainly not computationally difficult.", Roth & Franconeri, 2012, p. 7, see also Franconeri et al., 2012; Holcombe, Linares, & Vaziri-Pashkam, 2011; Yuan, Uttal, & Franconeri, 2016; see also Conder et al., 2017, who found neuronal activity in the superior parietal lobule during spatial language processing that was linked to shifts of attention by Molenberghs, Mesulam, Peeters, & Vandenberghe, 2007; but see also Hayworth, Lescroart, & Biederman, 2011 who argue against a serial interpretation of spatial relation processing).

In addition, spatial language processing research has focused on further aspects of spatial language processing (e.g., how functional and geometrical aspects of spatially related objects affect spatial language use, Carlson, Regier, Lopez, & Corrigan, 2006; Coventry et al., 2010; Coventry, Prat Sala, & Richards, 2001; Hörberg, 2008; Kluth & Schultheis, 2014; Landau, 2017). This appears of interest since the geometric properties of objects might themselves interact with mechanisms of spatial language processing and be relevant for related computational modeling.

In this article – following the demand of more "computational simulations" by Anderson et al. (2011, p. 188) – we are investigating the implications of this reversed shift (i.e., from the LO to the RO) for the AVS model. We do this by discussing a recently proposed model with a reversed vector sum (the rAVS model, Kluth, Burigo, & Knoeferle, 2017), empirically testing predictions arising from simulations of

[1] Note however that Lipinski et al. (2012); Richter et al. (2014, 2016, 2017) did not focus on this directionality of attention (which might be reversable in their models) but on capturing spatial language behavior with neuronally plausible mechanisms.

the rAVS model and assessing both models (AVS and rAVS) with state-of-the-art methods for model comparison (Navarro, Pitt, & Myung, 2004; Pitt, Kim, Navarro, & Myung, 2006; Schultheis, Singhaniya, & Chaplot, 2013; Veksler, Myers, & Gluck, 2015). Furthermore, we relate the implementations of attention in the cognitive models to the relevant literature in order to further "explicate the various theoretical claims" (Anderson et al., 2011, p. 188). We also consider the role that geometric properties of objects play for these models and in human behavior. The results of both – computational and empirical studies – provide insight into the role of (i) the directionality of attention and (ii) geometric properties of objects for spatial language processing and as such shed further light on the interaction between linguistic and visual processes.

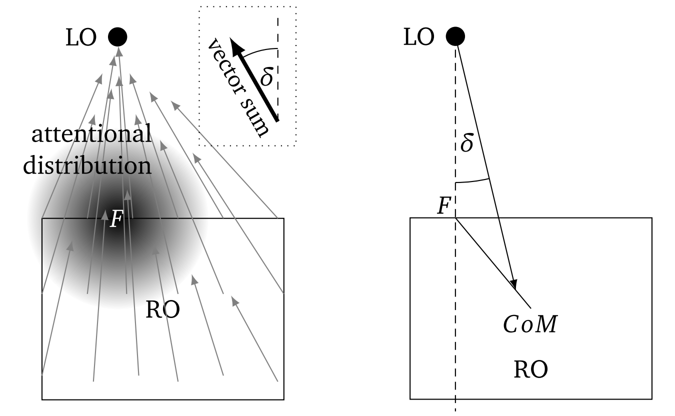
1.2. Overview of This Article

We structured our work of contrasting directionalities of attention in the following way. In Section 2.1, we start by introducing the AVS model (Regier & Carlson, 2001) as well as the reversed AVS model (Kluth, Burigo, & Knoeferle, 2017, a recent modification of the AVS model that reverses the direction of the vector sum). This is followed by a discussion about the role of the directed vector sum in terms of attentional shifts (Section 2.2) and the implications of using a directed vector sum to implement contrasting directionalities of attention (Section 2.3). Specifically, we identified two types of geometric shapes for which the two models predict somewhat different outcomes: rectangular ROs with different heights (testing the influence of 'relative distance', a geometric property assumed in the rAVS model) and asymmetrical ROs (testing different representations of the geometry of the RO in the two models). In Section 3, we present an empirical study investigating these specific model predictions. Using the collected empirical data and state-of-the-art methods for model comparison in Section 4, we aim to distinguish the two cognitive models – AVS and rAVS – and, in doing so, the assumptions about the directionality of attention they realize.

2. Models and Predictions

2.1. Models

Both, the Attentional Vector Sum (AVS) model (Regier & Carlson, 2001) and the reversed AVS (rAVS) model (Kluth, Burigo, & Knoeferle, 2017), have the same in- and output: Given the location and shape of an RO, the location of an LO, and a spatial preposition they compute an acceptability score for the sentence "The [LO] is [preposition] the [RO].". Both models also share the assumption that the relative location of the RO and the LO (as expressed by an angular deviation to a reference direction, canonical upright for *above*) contributes fundamentally to acceptability scores of projective spatial prepositions.



(a). Schema of the AVS model developed by Regier and Carlson (2001). F: attentional focus.

(b). Schema of the rAVS model developed by Kluth, Burigo, and Knoeferle (2017). CoM: center-of-mass.

Fig. 1. Schematized steps of (a) the AVS model and (b) the rAVS model. RO: reference object, LO: located object, δ: angular deviation from reference direction (dashed line).

2.1.1. AVS

To compute an acceptability score, the AVS model performs the following steps (see Figure 1a; for formulas see Regier & Carlson, 2001): First, it defines the location of an attentional focus F on the top of the RO (for *above*; F lies on corresponding parts of the RO for different prepositions, e.g., on the bottom of the RO for *below*). The attentional focus F lies at that point on top of the RO that is vertically aligned with the LO or closest to being so (see Figure 1a). Next, a distribution of attention is defined. The amount of attention is highest at the attentional focus F and decays exponentially (in Figure 1a a darker shading visualizes a higher amount of attention). Apart from the free model parameter λ, the strength of this decay is controlled by the distance of the LO to the RO. A close LO results in a narrow attentional distribution (attention drops off quickly within a small region around the attentional focus) whereas a distant LO results in a wide attentional distribution (attention drops off less quickly).

Next, a population of vectors is defined on the RO. Every vector points to the LO and is weighted by the amount of attention that was previously defined at this point. This gives vectors close to the focal point F a greater length (and hence importance) than vectors that are rooted farther away. All vectors are summed up to obtain a final direction. One of the two components that control the final outcome of the AVS model is then a linear mapping of the angular deviation δ of this final direction (compared to a reference direction, canonical upright for *above*) to an acceptability score: A high deviation leads to a low acceptability score whereas a low deviation leads to a high acceptability score. The *slope* and the *intercept* of the linear mapping function, are two additional free model parameters.

Acceptability scores are not, however, solely determined by the angular deviation. A second component of the AVS model (not depicted in Figure 1a) ranges from 0 to 1 and is multiplicatively combined with the outcome of the angular component. This other component uses the fourth free model parameter $highgain$. It identifies the vertical location of the LO relative to the RO, whereby the score from the angular component remains unchanged (the LO is well above the top of the RO), lowered slightly (the LO is close to the top of the RO), considerably (the LO is below the top of the RO), or drastically (the LO is below all points of the RO).

2.1.2. rAVS

The AVS model implements vector directionality (interpretable as attention direction, see Logan, 1995; Logan & Sadler, 1996) from the RO to the LO. But recent empirical findings about the real-time processing of spatial language (Burigo & Knoeferle, 2015) and the processing of spatial relations (Franconeri et al., 2012; Roth & Franconeri, 2012), suggest a reverse directionality, motivating the reversed AVS (rAVS) model Kluth, Burigo, & Knoeferle, 2017. The rAVS model implements an attentional shift from the LO to the RO (i.e., in reverse to the directionality in the AVS model).

The main computation steps in the rAVS model are similar to the AVS model. The rAVS model computes an angular deviation δ to a reference direction and maps it to an acceptability score. This score is adapted according to the vertical location of the LO. The computation of the angular deviation, however, was modified by changing the direction of the vectors. Instead of an attentional vector sum across the RO, the rAVS model defines an attentional vector sum on the LO (see Figure 1b for a visualization and Kluth, Burigo, & Knoeferle, 2017, for formulas). Due to the simplification of the LO as a single point in the AVS model and the desire to keep the rAVS model as close as possible to the AVS model, the vector sum in the rAVS model consists of only one single vector.

The choice of the vector destination is informed by previous observations. Regier (1996) and Regier and Carlson (2001) showed that the orientations of two imaginary lines connecting the LO with two important points of the RO can be used to predict human acceptability scores. These are (i) the ‘proximal orientation’ of the imaginary line that connects the LO with the proximal point – the point on the RO where RO and LO are closest to each other – and (ii) the ‘center-of-mass orientation’ of the imaginary line that connects the center-of-mass, CoM, of the RO with the LO (see Figure 1b). The more these orientations deviate from a reference direction, the lower people rated the appropriateness of the spatial preposition for the corresponding LO location (Regier & Carlson, 2001).

In the rAVS model, the vector from the LO to the RO always points on the line that connects the point F (the same as the attentional focus F in the AVS model[2]) with the point CoM (see Figure 1b). The exact vector destination is controlled by the distance between the LO and the RO. An LO with large distance from the RO yields a vector pointing close to the CoM of the RO. The closer the LO gets, the more the vector points towards point F. That is, the rAVS model considers the distance between the LO and the RO to weight the importance of the proximal and the center-of-mass orientations.

More precisely, the rAVS model uses the *relative* distance which is defined as the absolute distance divided by the dimension of the RO:

$$d_{rel.}(LO,RO) = \frac{|LO,P|_x}{RO_{width}} + \frac{|LO,P|_y}{RO_{height}} \qquad (1)$$

Here, $|LO,P|_x$ describes the horizontal component of the line connecting the LO with the proximal point P (see footnote 2); $|LO,P|_y$ describes the respective vertical component. As an example, consider the two RO-LO configurations shown in Figure 2a. Both LOs are, say, 5 cm away from the ROs. The first RO has a height of, say, 5 cm while the second RO has a height of, say, 30 cm. The relative distance in the first case would be 1 (5 divided by 5) and reduces in the second case to $\frac{1}{6}$ (5 divided by 30). Below, we present the results of an empirical test designed to detect whether *relative* distance affects human acceptability ratings.

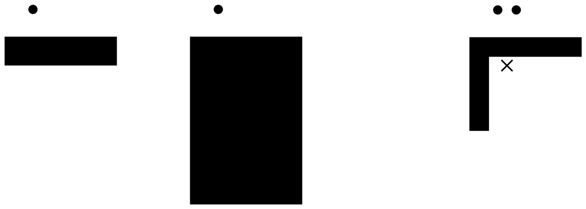
2.2. Vector Sum as an Attentional Mechanism

Let us first clarify the notion of “attention” in the models (Fernandez-Duque & Johnson, 1999). Regier and Carlson (2001) motivate their implementation of attention in the AVS model via a spotlight metaphor by calling the attentional distribution in the AVS model an “attentional beam” (Regier & Carlson, 2001, p. 277–278). Moreover, they refer to Logan (1994, 1995) who developed a theory of the apprehension of spatial relations and remark “that in several neural subsystems, overall direction is represented as the *vector sum* of a set of constituent directions” (p. 277, emphasis in the original, relevant references cited: Georgopoulos et al., 1986; Lee et al., 1988; Wilson & Kim, 1994). The explicit conceptualization of the vector sum in terms of attention, however, remains unclear.

We argue that the vector sum in the AVS and the rAVS model could be viewed as representing a directed movement of attention different from (but interacting with) the “spotlight-like” distribution of attention in the models. The attentional distribution selects one of the two objects of a spatial relation; the directed vector sum is related to where the “attentional spotlight” should move to next. This view fits well with the theory of Logan (1994, 1995); Logan and Sadler (1996) in which processing a spatial relation starts with “spatially indexing the arguments of the relation” (Logan, 1994, p. 1015, where spatial indices are theorized to be pre-attentive, Pylyshyn, 1989, 2001) and at a later stage “the viewer’s attention should move from the reference object to the located object” (Logan & Sadler, 1996, p. 499).

Interpreting the vector sum as an attentional movement echoes the sequential shift account of Franconeri et al. (2012).

[2] Note that attentional focus F in the AVS model does not always coinicide with the proximal point P. For instance, the closest point P for LOs to the right of an RO is located on the right side of the RO. The point F used in the AVS model and the rAVS model, however, is always located on the top of the RO (for *above*). In the rAVS model the proximal point P is additionally used for the computation of the relative distance, see eq. 1.



(a). ROs and example LOs used to test the role of relative distance.

(b). RO and example LOs used to test the influence of asymmetrical ROs. (× = center-of-mass)

Fig. 2. ROs and example LOs used as input for the PSP analysis.

While Franconeri et al. (2012) mainly focus on how attention spatially relates *two* objects, the AVS model was primarily concerned with the question how attention influences the processing of geometric properties of a *single* object of a spatial relation (the RO). For this, Regier and Carlson (2001) identified the vector sum as an adequate mechanism. The geometric properties of the RO affect the specific orientation of AVS's final vector direction and thus the outcome of the model. However, they do not affect the directionality of attention (i.e., whether the RO or the LO is selected first or second).

In putting less emphasis on the attentional distribution, the rAVS highlights the role of the directed vector sum. Specifically, the rAVS model still assumes a vector sum but roots it on the LO instead of on the RO. Doing so requires the rAVS model to find a different way of representing the geometry of the RO, as its vector sum now accounts for the geometry of the LO. Thus, the specific mechanisms of the rAVS model are implications of selecting the LO first and implementing a movement of attention from the LO to the RO (see Kluth, in preparation, for an evaluation of several different mechanisms that perform worse than the rAVS model).

The explicit use of relative distance is a specific mechanism that distinguishes the rAVS model from the AVS model. In particular, the rAVS model assumes that relative distance weights the influence of the proximal and the center-of-mass orientations on spatial language acceptability scores. Another distinguishing feature is the different representation of the geometry of the RO in the two models.

2.3. Assumptions and Predictions

On the *existing* data from Regier and Carlson (2001), Kluth, Burigo, and Knoeferle (2017) showed that both the AVS model and the rAVS model perform equally well. This is why we designed two kinds of geometric shapes for which the two models appear to predict different acceptability ratings. The first test case concerns RO-LO configurations that differ in relative distance (Figure 2a) and the second test case concerns asymmetrical ROs (Figure 2b). We first discuss these two test cases and corresponding model predictions followed by the associated empirical study on human participants (Section 3).

2.3.1. Relative Distance

rAVS. The rAVS model explicitly uses the *relative* distance between the LO and the RO for its computation. An LO relatively close to the RO results in a vector closer to the proximal point which in turn leads to a lower angular deviation and therefore to a higher acceptability score. An LO that is relatively far away from the RO, on the other hand, is rated lower by the rAVS model since the vector points more to the center-of-mass of the RO and thus a greater angular deviation emerges (i.e., a small relative distance leads to higher importance of the proximal orientation compared to the center-of-mass orientation whereas a large relative distance shifts this importance in favor of the center-of-mass orientation). Using this mechanism (averaging proximal and center-of-mass orientation using relative distance), the rAVS model successfully accounts for the data from Regier and Carlson (2001). However, whether this proposed mechanism is reflected in human behavior has not yet been tested. Thus, a prediction based on this mechanism serves as a test case for the implemented shift of attention from the LO to the RO.

The relative distance is computed from two sources: The absolute distance and the dimensions of the RO (see equation 1). Since there is already evidence for an effect of absolute distance on acceptability scores (e.g., Regier & Carlson, 2001), we only changed the dimensions of the RO to obtain stimuli with different relative distances. Consider the displays shown in Figure 2a. Here, the rAVS model rates LOs above the tall rectangle higher compared to LOs above the thin rectangle because the LOs above the tall rectangle are relatively closer than the ones above the thin rectangle and less relative distance leads to higher acceptability ratings (due to the greater importance of the non-deviating proximal orientation).

AVS. We used the same RO-LO configurations shown in Figure 2a to investigate the role of relative distance in the AVS model. The main component of the AVS model is its vector sum weighted by an attentional distribution. Since it is the *absolute* distance (which is the same for both configurations in Figure 2a) that influences the width of the attentional distribution, the attentional distribution is equal for both rectangles. What changes with rectangle height is the number of vectors in the vector sum. This is because Regier and Carlson (2001, p. 277) defined one vector "at each point of the landmark [RO]". Although this definition does not specify what measure should be used for a point in the RO (pixel, centimeters, . . .), it results in more points and thus more vectors for larger compared with smaller ROs. Compared to the thin rectangle, the tall rectangle has a greater area and thus more vectors. Note that the vector sum computed for the thin rectangle is completely contained in the vector sum for the tall rectangle (the upper part of the tall rectangle) such that only the additional vectors for the tall rectangle could change the final direction of the vector sum and its associated acceptability score.

Since the averaging mechanism in the AVS model is influenced by its free parameters, assessing the influence of these additional vectors is difficult. Depending on the values of the free parameters, the AVS model can generate either higher acceptability scores for the tall vs. the thin rectangle (i.e., like the rAVS model) or equal acceptability scores for both. In order to assess the capability of the AVS model, we applied the Parameter Space Partitioning (PSP, Pitt et al., 2006) algorithm, which reports all qualitative predictions that the model is able to compute. Before presenting the results of the PSP analysis, let us consider a further test case.

2.3.2. Asymmetrical Reference Objects

rAVS. Our second test case concerns asymmetrical ROs (Figure 2b). As part of implementing a shift of attention from the LO to the RO, the rAVS model relies on imaginary lines connecting the proximal points with the center-of-mass (symbol ×) of the RO. For the LOs shown in Figure 2b, these two lines are mirrored versions of each other since we placed the LOs symmetrically with respect to the center-of-mass. This means that the two deviations of rAVS's vectors only differ in their sign (same deviation either to the left or to the right). Thus, the rAVS model generates the exact same acceptability scores for both LOs. This is in line with previous research (Regier, 1996; Regier & Carlson, 2001) predicting equal acceptability scores for LOs with equal proximal and center-of-mass orientations.

AVS. The AVS model uses the center-of-mass orientation only implicitly in its weighted vector sum. It computes the center-of-mass orientation if the attentional distribution is of uniform strength (Regier & Carlson, 2001). For almost[3] all other attentional distributions, the final direction of the vector sum is harder to grasp due to the interplay of a weighted population of vectors controlled by free parameters. Arguably, however, the AVS model can be interpreted as predicting a higher acceptability score for the left LO in Figure 2b compared to the right LO. This is because the downward oriented "leg" on the left side of the L-shaped RO is populated by vectors but – due to the cavity of the RO – no vectors are rooted on the corresponding location on the right side of the RO. This asymmetry in the vector sum might lead to higher acceptability scores for the left LO (closer to the edge of the RO and with more mass directly below it) compared to the right LO.

2.3.3. Parameter Space Partitioning

We applied the method Parameter Space Partitioning (PSP, Pitt et al., 2006) that quantifies the range of qualitatively different model predictions (see Appendix B.2 for details). For the rAVS model, the PSP analysis confirmed our "intuitive" model predictions for a large range of different model parameters: lower ratings for LOs above the thin rectangle compared to the tall rectangle and equal ratings for the LOs to either side of the center-of-mass of the asymmetrical RO. The rAVS model generates only one additional data pattern (with a small value of its parameter α, equal ratings for different relative distances).

For the AVS model, the PSP analysis show a more diverse pattern of model predictions. For the asymmetrical RO, the AVS model does not generate our "intuitive" prediction of a higher rating for the left LO compared to the right LO (Figure 2b). Rather, the model predicts either no difference or a higher rating for the right LO. For relative distance variation, the PSP analysis revealed that the AVS model predicts either (i) no different ratings for LOs above the thin versus the tall rectangle or (ii) lower ratings for LOs above the thin rectangle compared to the tall rectangle. In summary, the AVS predictions vary more than the rAVS predictions but surprisingly the predictions of both models are the same for a considerable range of parameter settings.

3. Empirical Study

We examined to what extent humans follow the PSP predictions by conducting an empirical acceptability rating study. Additional (eye-movement) data that are less central to the main argument are presented in Appendix A.3.

3.1. Materials and Procedure

Materials. We tested all of the geometric shapes used in the PSP analysis plus five extra ROs to generalize the predictions and to collect more data[4]. We placed 28 LOs above and below each RO. 4 LOs out of these 28 LOs were placed at the same height as or slightly below/above the top/bottom of the RO. The remaining 24 LOs were arranged in a grid with 3 rows and 8 columns (see Figures 4, 5, 6, and 8 for visualizations of the ROs and the placement grid including row and column coding). For each of the 28 LOs above each of the 8 ROs (rows R1–R5), participants had to rate how well the German sentence "Der Punkt ist über dem Objekt." ("The dot is above the object.") described the depicted layout. For all LOs below the ROs (rows R6–R10), the accompanying sentence was "Der Punkt ist unter dem Objekt." ("The dot is below the object."). In order to keep the surface of the RO that faces the LO constantly flat in all conditions, we horizontally mirrored the L and mL objects for the 28 LOs below these ROs (see Figure 8 on page 12).

Participants saw each RO-LO combination exactly once and only one RO and one LO were present on a single trial. The center-of-mass of the RO was always centered on-screen. We placed the LOs relative to the borders of the ROs such that their absolute distances to the corresponding RO were equal for all ROs. Taken together, this rating study consisted of 8 ROs × 28 LOs × 2 prepositions = 448 items. Participants sat in front of a computer monitor (22 inches, 1680 × 1050 pixel) at a distance of approximately 80 cm. Their right eye was tracked with a desktop mounted eye tracking system (EyeLink 1000, SR Research) using a chinrest. We programmed the

[3] The other extreme case of an attentional distribution (i.e., 1 at the focal point and 0 at all other points) yields the orientation of a line connecting the LO with the focal point (this is the proximal orientation if focal point F and proximal point P coincide).

[4] For the relative distance test case, we added a thick rectangle and a square (see Figure 4 on page 9). For the asymmetrical ROs test case, we added a C-shaped RO as well as mirrored versions of the C- and the L-shaped ROs to balance potential left-right biases (named mC and mL, respectively; see Figure 6 on page 11 and Figure 8 on page 12).

experiment in “Experiment Builder” (version 1.10.1025, SR Research). The study was approved by the Bielefeld University ethics committee (2015-126).

Procedure. We recruited 34 participants (19 females; 18–34 years, M=23.79). Most of the participants were either students at the Bielefeld University or the University of Applied Science Bielefeld. Each participant received 6 € for participation. The study took approximately 45 minutes. The participants answered a general questionnaire and were asked to rate each picture-sentence pair they would see for how well the sentence matched the picture (using keys 1–9 above the letters on a standard keyboard). Here, 1 meant “The sentence does not describe the picture at all.” and 9 meant “The sentence describes the picture very well.”. Participants were encouraged to use the entire rating range. After eye tracker calibration, participants rated four practice trials (with different, non-critical ROs) and then all 448 items in a pseudo-random order (items were randomized but the same RO did not appear twice in a row). Participants were presented with the sentence “Der Punkt ist über/unter dem Objekt” (“The dot is above/below the object”; only one preposition shown in one trial) and pressed the space bar once they had read it. Then, one RO and one LO appeared on the screen, until participants gave their rating. RT was measured from the onset of the picture until the rating response.

3.2. Results

Method. All following data analyses were conducted using the Bayesian framework. There is growing consensus that the classical Null Hypothesis Significance Testing (NHST) framework focusing on the significance of an effect given a corresponding $p < 0.05$ has severe flaws (e.g., Dienes, 2011; Gigerenzer, 2004; Kruschke, 2013; Lindley, 1993; Wagenmakers, 2007; Wagenmakers et al., 2018). Bayesian data analysis overcomes most of the problems of the NHST (e.g., Etz, Gronau, Dablander, Edelsbrunner, & Baribault, 2018).

In the Bayesian framework, we used multilevel regression models to describe our data. Since our study has a repeated measurement design, we included subjects as a group-level term to account for intersubject variability. For the analysis of the acceptability ratings, we applied ordinal regressions (as the predicted variable rating is ordered and discrete) collapsing across *über* (*above*) and *unter* (*below*; see Appendix A.1 for additional analyses). This type of regression uses a metric variable underlying the ratings. The slope coefficient of an ordinal regression gives us information about how the latent metric variable changes with respect to the values of the predictor(s). The larger the absolute value of slope, the higher the change in ratings. However, one cannot directly interpret the value of the slope on the scale of ratings.

We report the estimated values of the regression parameters of interest (means of the corresponding posterior distributions) together with their 95% credible intervals (CI) that contain 95% of the probability density of the posterior distributions (and thus are a measure of the uncertainty of the estimation). We ran all following analyses using R (R Core Team, 2016) with the R package `brms` (Bürkner, 2017). If not specified otherwise, we used the default prior distributions (designed to be non-informative) provided by the `brms` package. Regression models with manually specified prior distributions resulted in the same qualitative output (existence and direction of effects) as the same regression models with `brms`’s default priors.

We sampled from the posterior distributions with four chains (with each providing 1000 warmup samples and 3000 post-warmup posterior samples; in total 12000 post-warmup samples) and verified that all models had a sufficient number of effective samples. We checked that all chains converged using the potential scale reduction statistic $\hat{R}$ (Gelman & Rubin, 1992). Moreover, we performed visual posterior predictive checks with the help of the R package `bayesplot` (Gabry, 2017). Where applicable, we compared different regression models using the leave-one-out cross-validation method (LOO, Gelman, Hwang, & Vehtari, 2014; Vehtari, Gelman, & Gabry, 2017) and the widely applicable information criterion (or Watanabe-Akaike information criterion, WAIC, Vehtari et al., 2017; Watanabe, 2010; both the LOO and the WAIC are goodness-of-fit measures that are adjusted for over-fitting by considering the effective number of model parameters, Gelman et al., 2014; Vehtari et al., 2017).

We used the software “Data Viewer” (version 1.11.900, SR Research) to generate (i) a trial report containing the acceptability ratings and (ii) a fixation report used to analyze participant’s eye movements (for the analysis of the eye movement data see Appendix A.3). All empirical data files, the fits of the Bayesian models as R data files, and R source code files to reproduce the fits of the Bayesian models are available in the data publication that also includes an implementation of the cognitive models ([dataset]Kluth, 2018).

3.2.1. Acceptability Ratings

Figures 4, 5, 6, and 8 present visualizations of the empirical ratings: Each rhombus represents one individual rating (the darker the color, the higher the rating). The figures also include row and column numbers which will be used in the analysis to refer to subsets of LOs. Note that for Figure 6 *über* and *unter* (*above* and *below*) ratings are depicted in the same image although participants did not rate *über* and *unter* for every LO (see Appendix A.1 for further details).

Relative Distance. In the PSP analysis, the AVS and the rAVS model both predicted higher ratings for LOs above the tall rectangle compared to LOs above the thin rectangle. To analyze whether our participants followed this prediction, we specified an ordinal regression that predicted rating by the type of rectangle (thin, thick, square, tall). As priors for each of the slope parameters of the regression model we chose Gaussian distributions with $\mu = 0.5$ and a large standard deviation $\sigma = 1.5$. The standard deviation assign relatively large probabilities for values ≤ 0.0, i.e., they allow for a null effect and also for a negative effect (higher rectangle leads to lower ratings). The positive mean $\mu = 0.5$ of the prior distributions

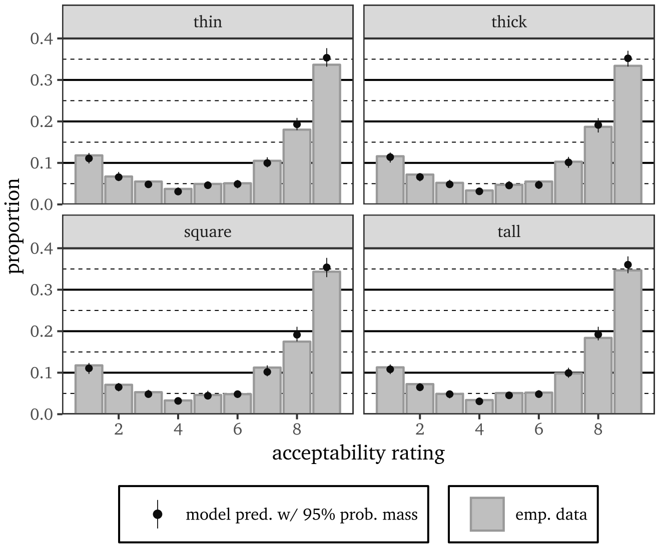


Fig. 3. Posterior predictive check for the regression model that predicted rating as a function of rectangular RO. Computed with 100 samples from the posterior distribution.

reflects our "trust" in the positive predicted effect of the AVS and rAVS models.

In contrast to the AVS and rAVS prediction, our analysis provides no evidence for different rating patterns across rectangular ROs with different heights (see Figure 3). Although all mean estimates of the regression slopes (comparing ratings for LOs above the thin rectangle vs. the three taller rectangles) were positive, their 95% credible intervals include 0.0 ($\beta_{thick} = 0.01$, 95% CI $[-0.11, 0.12]$; $\beta_{square} = 0.02$, 95% CI $[-0.09, 0.14]$; $\beta_{tall} = 0.04$, 95% CI $[-0.08, 0.15]$).

The rAVS model uses relative distance to weight the importance of the proximal and the center-of-mass orientation. From the shorter to the taller rectangles, relative distance reduces – so the rAVS model assumes that (i) the importance of the proximal orientation increases while (ii) the importance of the center-of-mass orientation decreases with increasing height of the rectangles. We accordingly predicted the ratings via relative distance, center-of-mass orientation and proximal orientation, allowing full interactions between these three predictors in an ordinal regression (see Appendix A.1.1 for details). In that analysis, relative distance modulated the influence of proximal orientation but different from the rAVS prediction: Higher relative distance *strengthened* the influence of proximal orientation (see different slopes in Figure A.10a on page 20). It further amplified a reversed effect of the center-of-mass orientation for high values of proximal orientation (i.e., higher center-of-mass orientation resulted in higher instead of lower ratings, see right subplot of Figure A.10b). These findings go against the specific mechanism implemented in the rAVS model.

Although the AVS model does not explicitly formulate the influence of relative distance on center-of-mass orientation and proximal orientation, its vector sum mechanism resembles the mechanism of the rAVS model (close LOs result in small attentional widths which in turn approximate the proximal orientation in contrast to the center-of-mass orientation). This suggests that the AVS model – just like the rAVS model – cannot fully accommodate the rating data.

Asymmetrical Objects. For the asymmetrical objects, both the AVS and the rAVS model predict equal ratings for LOs that have the same center-of-mass orientation (e.g., LOs in column C3 compared to LOs in column C4 for the C or the L, see Figures 6 or 8) and different ratings when center-of-mass orientation differs. The PSP analysis further revealed that the AVS model predicts higher ratings for LOs above (vs. not above) the cavity of an asymmetrical RO (see Section 2.3.3 or Appendix B.2). This calls the claimed effect of the center-of-mass orientation (Regier, 1996; Regier & Carlson, 2001) into question (the center-of-mass orientations are equal for the two sets of LOs we compared, one would not expect different ratings).

An ordinal regression model predicted ratings based on the subsets used for the PSP analysis ("mass" subset: ratings for LOs in columns C2 and C3 for the L and the C and columns C6 and C7 for the mL and the mC; "cavity" subset: ratings for the LOs in columns C4 and C5 for all ROs, see Figures 6 and 8). Based on the effect of the center-of-mass orientation (which is equal for both subsets) and constant proximal orientation for all LOs, we used a prior that reflects our expectation of finding no difference in ratings (Gaussian distribution with $\mu = 0.0$ and $\sigma = 0.1$ as prior on the slope coefficient).

The posterior distribution of the slope coefficient, however, was credibly different from 0.0 with a mean estimate of 0.84 and a 95% credible interval from 0.71 to 0.97. This provides evidence that – despite equal center-of-mass orientations – ratings were higher for LOs in the "cavity" subset compared to the "mass" subset (8.8% more probability for the rating 9 if the LO is in the "cavity" versus "mass" subset). A second model, using the default, non-informative prior from the `brms` package replicated these results (14.7% higher probability for rating 9 if LO is in subset "cavity" vs. "mass"; slope coefficient: 1.46, 95% CI: [1.29, 1.63]). This second model fitted the data better than the model with the prior that emphasized our null-effect expectation (LOO for model with restrictive prior: 5680.88; for model with default prior: 5631.44; lower LOO is better, see Vehtari et al., 2017). Figure 7 plots the predictions of the regression model with the default prior alongside with the empirical data.

Overall, the model predicts the empirical data well but the LOs in the two subsets were rated differently. Figure 7 illustrates that LOs in the "cavity" subset received considerable more 9s in contrast to LOs in the "mass" subset (lower values). These results conflict with (r)AVS prediction of equal ratings for LOs with same center-of-mass orientation. However, they confirm the PSP-prediction from the AVS model (higher ratings for LOs in the "cavity" vs. "mass" subset).[5]

[5] Note, however, that the strength of this prediction from the AVS – as measured in terms of covered volume in the parameter space – is considerably

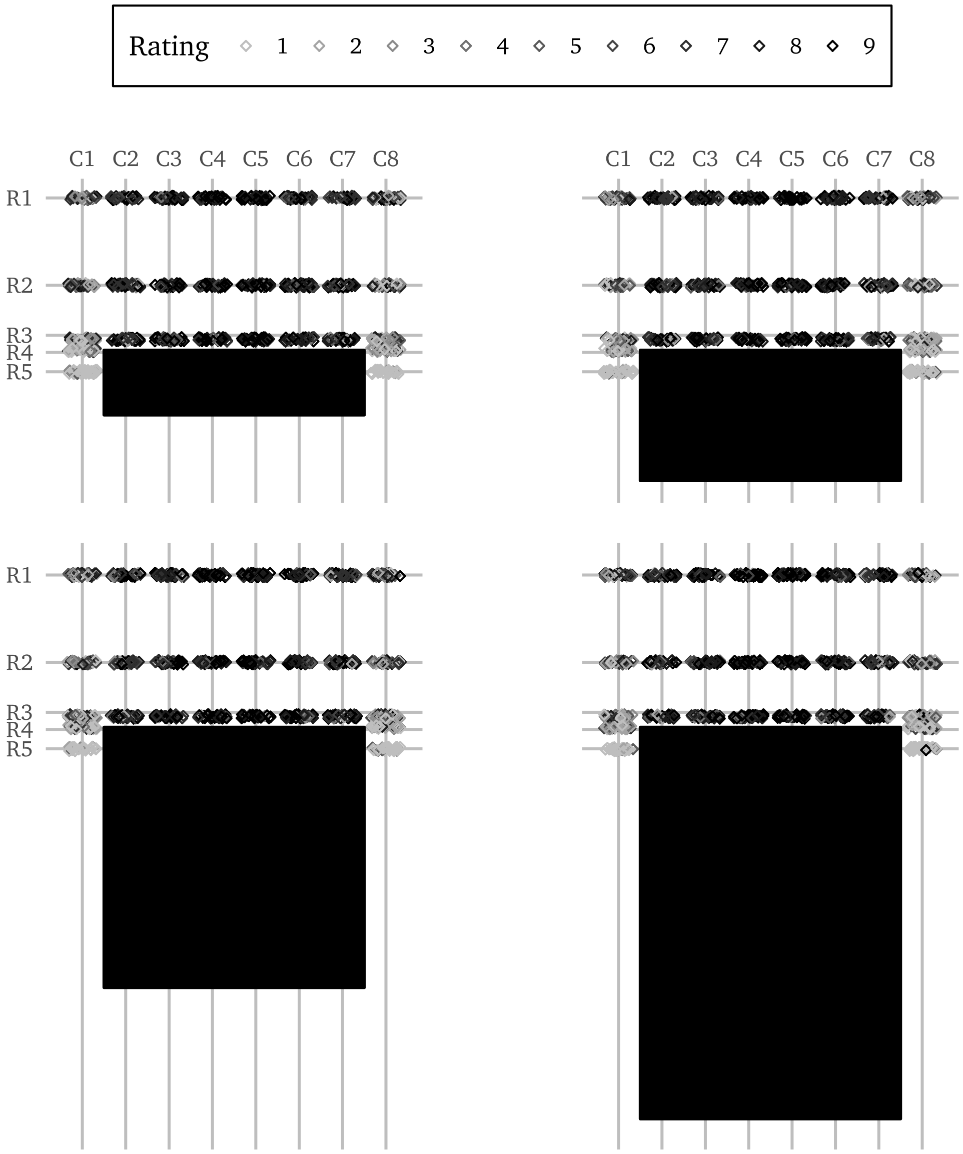


Fig. 4. Visualization of individual *über* (*above*) acceptability ratings for LOs above the thin, the thick, the square, and the tall rectangle. LOs (not depicted) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). Only one RO and one LO was visible at a time.

The LOs that were rated higher in the previous comparison are all more central above the center-of-*object* of the RO[6] (compared to the lower-rated LOs). Using (post-hoc) the center-of-object orientation instead of the center-of-mass orientation to explain human acceptability ratings can account for the pattern of results that we found (see Appendix A.2). This is in contrast with the idea that humans use the center-of-*mass* of the RO as a base for their acceptability ratings (as proposed by Regier, 1996; Regier & Carlson, 2001).[7]

To contrast the explanatory power of the two predictors

larger if the difference threshold for equality of model-generated ratings is $t_e = 0.1$ (more than 65%) compared with when it is $t_e = 0.5$ (less than 4%; see Figure B.13 on page 26 in the Appendix). Accordingly, the AVS model predicted that the difference of ratings for the two subsets should be rather small (less than $t_e = 0.1$). This suggests that the empirical results reflect a clearer difference in ratings for the LOs in the "cavity" versus the "mass" subset than predicted by the AVS model.

[6] We define the center-of-object as the point that lies in the center of the bounding box of the RO (the smallest rectangle containing all points of the RO, see dashed lines in Figures 6 and 8). More formally, this corresponds to the point $CoO(x,y) = \left(RO_{x0} + \frac{RO_{width}}{2}, RO_{y0} + \frac{RO_{height}}{2}\right)$, where RO_{x0} is the leftmost point of the RO and RO_{y0} is the point of the RO with the lowest y-coordinate (y-axis growing from bottom to top). For the rectangular ROs, this point is the same as the center-of-mass; for the asymmetrical ROs, the center-of-object is different from the center-of-mass. Figures 6 and 8 mark the location of the center-of-mass with the symbol × and the location of the center-of-object with the symbol ∘.

[7] Note that this conflicts with the results from experiment 4 conducted by Regier and Carlson (2001). However, in their experiment only 8 LOs above 2 different ROs were tested and the advantage of the center-of-mass over the center-of-object was quite small (but significant). A related model by Lovett and Forbus (2009) also failed to replicate the effect in this particular experiment. We speculate that whether the side of the RO that faces the LO is flat (or not) qualifies the different influences of the center-of-mass

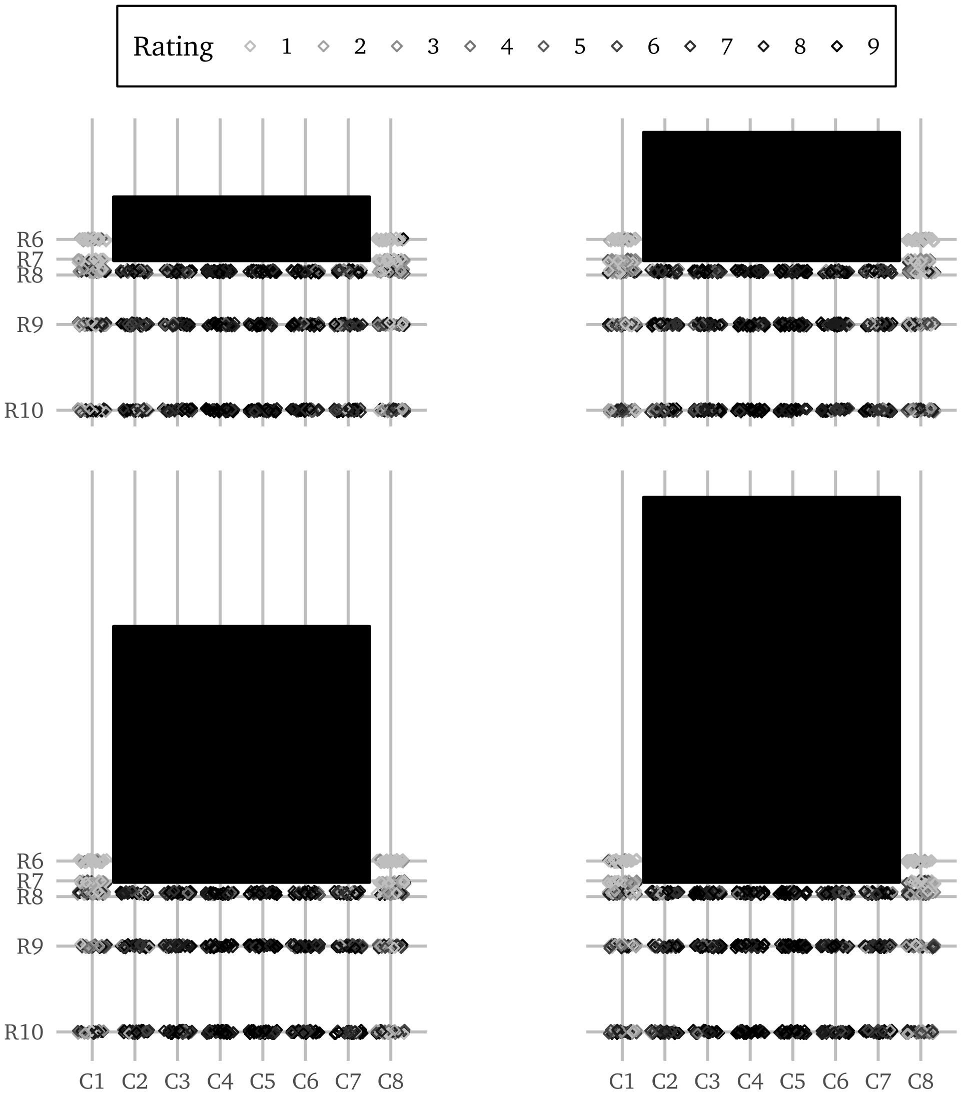


Fig. 5. Visualization of individual *unter* (*below*) acceptability ratings for LOs below the thin, the thick, the square, and the tall rectangle. LOs (not depicted) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). Only one RO and one LO was visible at a time.

center-of-mass orientation and center-of-object orientation we finally ran two ordinal regressions using only one of the two predictors (in radian notation and with default, non-informative priors). Both of these models resulted in a credible effect of the corresponding predictor ($\beta_{CoM} = -4.58$, 95% CI $[-4.73, -4.42]$; $\beta_{CoO} = -7.24$, 95% CI $[-7.46, -7.02]$). The center-of-object orientation, however, had a greater effect on the ratings than the center-of-mass orientation as revealed by the magnitude of the regression coefficients. Moreover, the model that used the center-of-object orientation as predictor also fitted the data better according to the LOO method (center-of-mass model LOO: 23 235.51, center-of-object model LOO: 21 175.10).

3.3. *Discussion*

In summary, the analyses of the ratings revealed an effect of relative distance but different from what the models had predicted. Both the AVS and the rAVS model – despite different directionalities of attention – predicted higher ratings for LOs above taller rectangles compared to ratings for LOs above shorter rectangles. This is what one would also expect when reasoning only with the center-of-mass orientation. The AVS model is also capable of computing no difference in ratings, which is the null-effect that we found. We showed in our analysis that the higher the relative distance is, the higher the influence of the proximal orientation becomes. In addition, for high values of proximal orientation, higher relative distance correlated with a stronger reversed effect of center-of-mass orientation (i.e., higher center-of-mass orientation resulted in higher ratings) than lower relative distance. Thus, our empirical results provide some evidence against the

or the center-of-object orientation (compare also Regier & Carlson, 2001, experiment 5 with our L shaped ROs). Since we only used ROs with a flat top/bottom, more studies are needed to provide additional evidence.

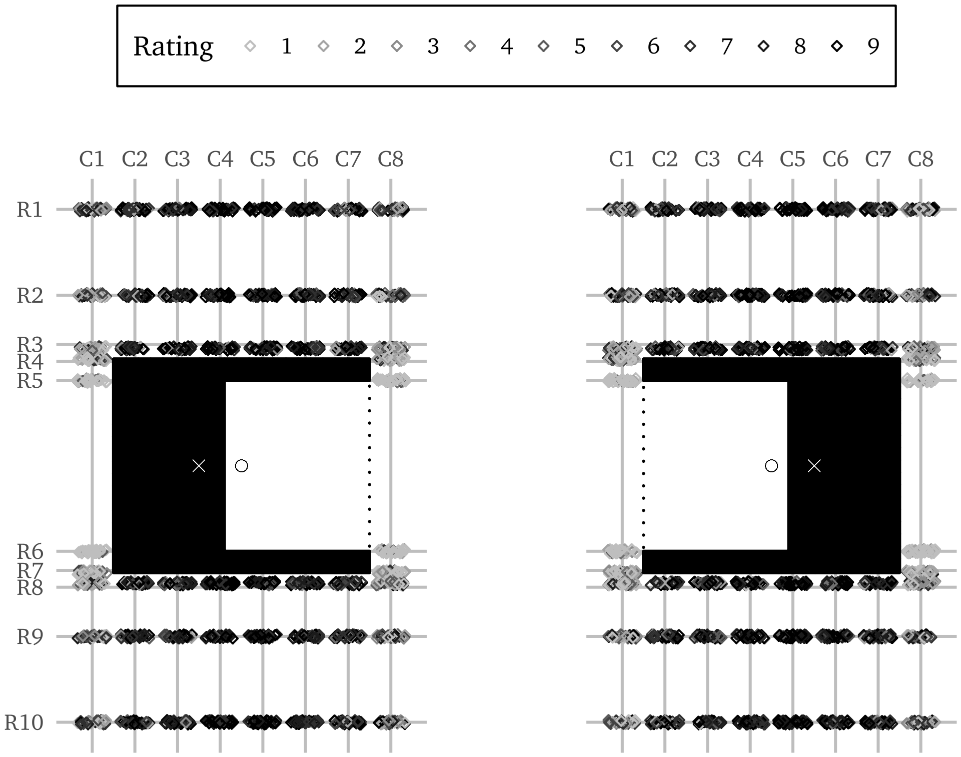


Fig. 6. Visualization of individual *über* (*above*) and *unter* (*below*) acceptability ratings for LOs around the asymmetrical C and mC ROs. LOs (not depicted) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs in rows R1–R5 were presented with *über* (above), LOs in rows R6–R10 were presented with *unter* (below). Only one RO and one LO was visible at a time. For each RO: Dashed line is the bounding box, × is the center-of-mass, ∘ is the center-of-object. Neither of the centers nor the bounding box were visible to the participants.

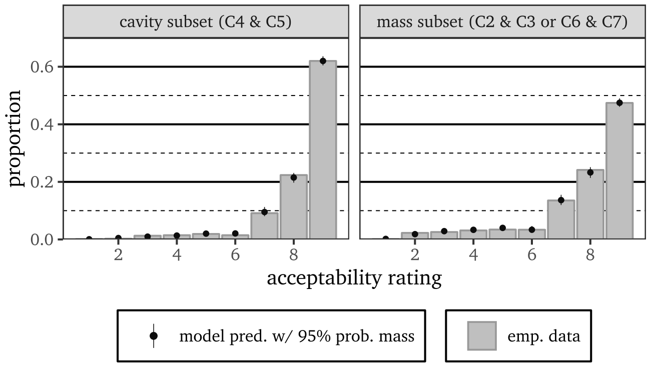


Fig. 7. Posterior predictive check for the regression model that predicted rating from the location of the LO (either in the "'cavity" or "mass" subset). Computed with 100 samples from the posterior distribution.

rAVS use of relative distance to modulate influences of the center-of-mass and proximal orientations.

Our findings for the asymmetrical ROs deepen the knowledge of effects of geometry on spatial language acceptability scores. In line with previous research highlighting the importance of the center-of-mass orientation (Regier, 1996; Regier & Carlson, 2001), both the AVS and the rAVS model predicted equal ratings for two LOs placed with an equal center-of-mass orientation. Despite this, our participants reliably rated these two LOs differently. The LOs that were more central with respect to the center-of-object seemed to match a more prototypical use of *über* (*above*; and *unter*, *below*) suggesting that people use the center-of-object orientation instead of the center-of-mass orientation.

4. Model Simulations

Based on the empirical results, we introduce two modified versions of the cognitive models AVS and rAVS (Section 4.1). These integrate the unexpected finding of the seemingly greater importance of the center-of-object compared to the center-of-mass by using the center-of-object.[8] We applied several model comparison techniques that provide different perspectives on the implications of the implemented attentional shifts for the performance of all cognitive models. Specifically, we fitted the models to the empirical data

[8] Although the analysis of the effect of relative distance also revealed findings that neither the AVS nor the rAVS model can explain, we do not propose modifications based on this effect. The main reason for this is that we would need to adapt the mechanism of the interaction of proximal and center-of-mass orientation (because our empirical findings revealed that it is modulated by relative distance). However, this mechanism is central to both implemented directionalities of attention and, in addition, an adaption is not as straight-forward as changing the center-of-mass to the center-of-object. Moreover, since we would need to change a core part of the models, we would obtain substantially different models. This would complicate model comparison and further entail the evaluation of alternative mechanisms, a step that goes beyond the scope of the present article.

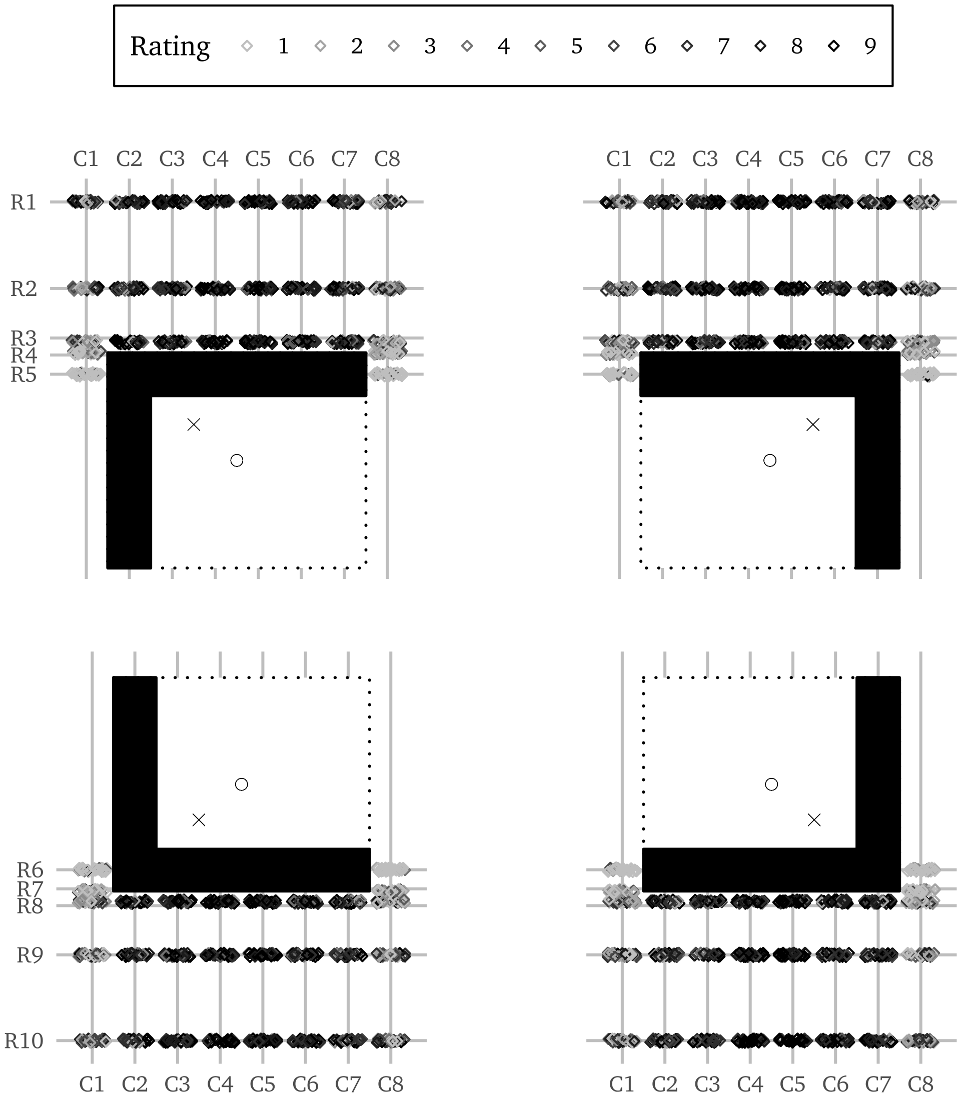


Fig. 8. Visualization of individual *über* (*above*) and *unter* (*below*) acceptability ratings for LOs around the asymmetrical L and mL ROs. LOs (not depicted) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs in rows R1–R5 were presented with *über* (above), LOs in rows R6–R10 were presented with *unter* (below). Only one RO and one LO was visible at a time. For each RO: Dashed line is the bounding box, × is the center-of-mass, ∘ is the center-of-object. Neither of the centers nor the bounding box were visible to the participants.

(Section 4.2), investigated their flexibility (Section 4.3), and analyzed the informativeness of the empirical data for distinguishing between the model assumptions about the directionality of the attentional shift (Section 4.4).

4.1. Implementing the Center-of-Object

AVS-BB. As noted by Regier and Carlson (2001), the AVS model computes the center-of-mass orientation for a uniform attentional distribution because all points of the RO are then equally weighted in the vector sum. Since the center-of-object is the center of the bounding box, we extended the vector sum to all points inside the bounding box to obtain the 'AVS bounding box' model (henceforth AVS-BB model).[9]

rAVS-CoO. In the rAVS model, the vector pointing from the LO to the RO points on a line that connects the center-of-mass with the point on top of the RO that is vertically aligned with the LO (see Figure 1b). In the here proposed 'rAVS center-of-object' – rAVS-CoO – model, this line connects the center-of-object (instead of the center-of-mass) with the point on top of the RO. All other steps in the rAVS-CoO model remain the same as in the rAVS model.

[9] Having defined the AVS-BB model in such way, we note that an asymmetrical RO with an $x \times y$ sized bounding box will be treated exactly the same as an $x \times y$ rectangle, which might be a problematic assumption for asymmetrical ROs with non-flat tops (e.g., used in exp. 5 by Regier & Carlson, 2001).

4.2. *Fitting the Models to the Data*

4.2.1. *Goodness-of-Fit*

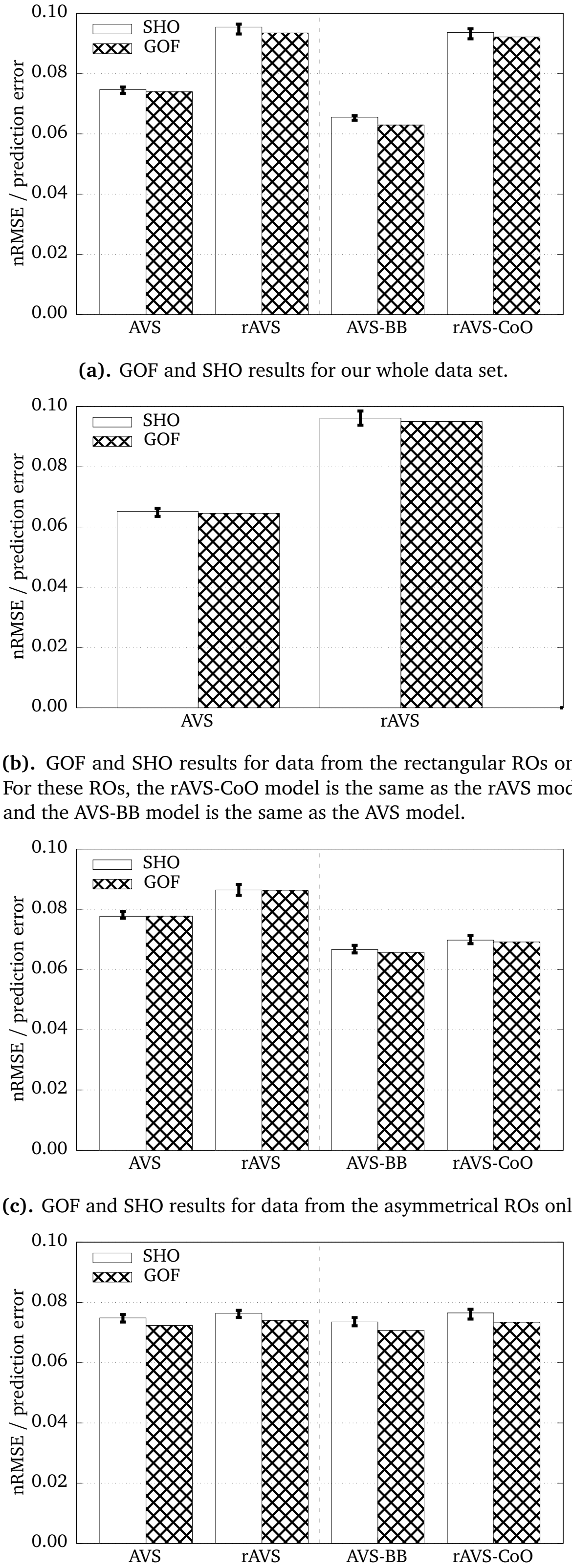
Method. As is common in the assessment of cognitive models on empirical data, we evaluate the goodness-of-fit (GOF) by minimizing the difference of model output to empirical data. A common measure of GOF is the Root Mean Square Error (RMSE). We additionally normalized the RMSE by dividing the RMSE with the rating range (obtaining the nRMSE: normalized RMSE) to be able to compare model fits across studies with different rating ranges (see Appendix B.1 for more details). We computed the GOF for our complete data set, data from the rectangular ROs only, data from the asymmetrical ROs only, and the complete data set from Regier and Carlson (2001)[10]. We computed the GOF for the AVS, the rAVS, the AVS-BB, and the rAVS-CoO model.

We implemented all models and the GOF computation (as well as all other model evaluation techniques presented later) in C++ using the CGAL library (The CGAL Project, 2015) and the GNU scientific library (GSL, Galassi et al., 2009). The documented source code is available under an open source license in [dataset]Kluth (2018).

Results. The GOF values for all models and subsets are shown as textured bars in Figure 9. These GOFs provide evidence for good model performance on all data sets (all nRMSE values < 0.1, worst possible nRMSE is 1.0). The AVS model had lower GOFs than the rAVS model for our complete data set (Figure 9a), as well as for the two subsets (Figure 9b and 9c). This difference was most pronounced for data from the rectangular ROs and less clear for data from the asymmetrical ROs. The GOFs for the complete data set were intermediate compared to the subsets.

Interestingly, the AVS-BB model and the rAVS-CoO model obtain considerably better GOF values for the rating data from the asymmetrical ROs[11] compared to the unmodified models (Figure 9c). This supports our suggestion that people rather use the center-of-object orientation instead of the center-of-mass orientation. If we compare the GOF of the AVS-BB model with the GOF of the rAVS-CoO model we see very little advantage for the AVS-BB model for the asymmetrical ROs (Figure 9c) but a pronounced advantage for the complete data set (Figure 9a). Apparently, the difficulties of the rAVS(-CoO) model to fit the data for the rectangular ROs (Figure 9b) is more strongly reflected in the GOF for the whole data set as is the better GOF of the rAVS-CoO for the asymmetrical ROs. On the data from Regier and Carlson (2001, Figure 9d), the AVS-BB and the rAVS-CoO model perform as well as the AVS and the rAVS model.

A good fit to empirical data is a necessary condition for a cognitive model to be considered an appropriate model of cognitive processes and representations. The models considered here all fulfill this condition. Complementary model



(a). GOF and SHO results for our whole data set.

(b). GOF and SHO results for data from the rectangular ROs only. For these ROs, the rAVS-CoO model is the same as the rAVS model and the AVS-BB model is the same as the AVS model.

(c). GOF and SHO results for data from the asymmetrical ROs only.

(d). GOF and SHO results for the whole data set from Regier and Carlson (2001)

Fig. 9. Goodness-of-fit (GOF) and Simple Hold-Out (SHO) results for (a)–(c) our data (collapsing across *über*, *above*, and *unter*, *below*) and (d) data from Regier and Carlson (2001). Error bars show bootstrapped 95% confidence intervals of the SHO median.

[10] We thank Terry Regier and Laura Carlson for sharing these data.

[11] For the rectangular ROs, the new modified models are the same models as the unmodified models (because the center-of-mass coincides with center-of-object for rectangles).

assessment techniques account for the generalizability of the models' output (Pitt & Myung, 2002), for instance by considering potential over-fitting of the models (better GOFs due to better fitting noise in the data). Schultheis et al. (2013) showed that the simple hold-out (SHO) method successfully controls for over-fitting which is why we applied it to our data and models.

4.2.2. Simple Hold-Out

Method. The SHO method is a cross-fitting method that computes a so-called prediction error. To do so, it randomly splits the data set into a training-set and a test-set. We used 70% of the data as training-set and 30% as test-set. The model's parameters are estimated on the training-set (i.e., the model is fitted to the data) and used to compute an nRMSE on the test-set. This nRMSE is called prediction error, because the data in the test-set is new to the model (but not new in an empirical sense). This process is done several times with the median of all prediction errors reported as the final outcome of the SHO method. We used 101 iterations to obtain a clearly defined median and also computed the 95% confidence intervals of this median using R with the `boot` package (Canty & Ripley, 2016).

Results. We plotted the results of the SHO method as white bars with 95% confidence intervals next to the GOFs in Figure 9. All SHO medians are close to the corresponding GOFs indicating a neglectable influence of over-fitting in the GOFs for any model. Thus, the SHO method confirms the general trend already discussed for the GOFs.

Taken together, the GOF and SHO results favor (a) the AVS model over the rAVS model (for our data, for the data from Regier & Carlson, 2001, both models perform equal as already shown in Kluth, Burigo, & Knoeferle, 2017) and (b) the modified models that incorporate the center-of-object orientation instead of the center-of-mass orientation over the unmodified models for the relevant ratings from the asymmetrical ROs.

4.3. Model Flexibility

GOF and SHO values report the performance of a model given a particular data set. While this is a valuable and important measurement to judge the quality of a model, it is not sufficient for a thorough model evaluation (e.g., Roberts & Pashler, 2000). Regardless of an empirical data set, it is of interest what a model can and what it cannot compute, as this gives information about how the model – here, the implemented assumption about the directionality of attention – constrains future empirical data. Among other methods, this model property can be measured with the PSP method that provides a list of all qualitative data patterns a model can generate (given a set of stimuli). We already applied the PSP method for the AVS and the rAVS model to generate the predictions – implications of implementing different directionalities of attention – for our empirical study (Section 2.3.3). However, given that the AVS-BB and the rAVS-CoO model better accounted for the empirical data – while still implementing contrasting shifts of attention –, we were also interested in analyzing their output possibilities using the PSP algorithm.

Our PSP analysis confirmed that the newly proposed models successfully accounted for the rating pattern found for the asymmetrical ROs while more work is needed to fully understand the role of relative distance for spatial language acceptability scores. In terms of the number of generated data patterns, the PSP analysis showed that the AVS-BB model generates 4 out of $3^2 = 9$ data patterns rendering it slightly more flexible than the rAVS-CoO model that only generates 3 patterns (for a more in-depth discussion of the PSP results see Appendix B.2.2). An additional Model Flexibility Analysis (MFA, Veksler et al., 2015, see Appendix B.3) that provides quantitative model flexibilities (instead of qualitative rankings like the PSP) revealed that (i) the AVS-BB model is less flexible than the AVS model and (ii) the rAVS-CoO model is less flexible than the rAVS model.

In summary, compared to the AVS and the rAVS model, the AVS-BB and the rAVS-CoO model (i) fit the empirical data better (GOF and SHO), (ii) generate data patterns that are closer to the empirical patterns (PSP) while (iii) being less flexible (MFA). This suggests superior performance of the two newly proposed models.

While the data from the asymmetrical ROs helped in distinguishing the AVS from the rAVS model, these data are possibly not informative enough to differentiate the more appropriate models (i.e., the AVS-BB and the rAVS-CoO model). This is why we applied a further "global" model analysis (the landscaping method, Navarro, Myung, Pitt, & Kim, 2003; Navarro et al., 2004) trying to distinguish between the AVS-BB and the rAVS-CoO model.

4.4. Landscaping

Method. The prime motivation for the development of the landscaping method was to provide a measure that helps to "assess [the] potential distinguishability [of competing models] and the informativeness of a data set in deciding between them" (Navarro et al., 2004, p. 48). Landscaping provides a measure of what is called *model mimicry* by Wagenmakers, Ratcliff, Gomez, and Iverson (2004): The ability of a model to account for data generated by another model. Each model should fit self-generated data quite well – without added noise this fit should be almost perfect. If, however, one model is also able to closely fit the data generated by another model, this model mimics the other model, i.e., this model is able to behave like the other model.

If the AVS-BB and the rAVS-CoO model are mimicking each other on our asymmetrical ROs, this means that despite their different assumptions and implementations regarding the directionality of attention, they are generating similar data for these stimuli – which could explain why we cannot distinguish between them (and their corresponding claims about attention). Alternatively, the models do not mimic each other, suggesting that perhaps the different implementations of the directionality of attention produce different model predictions. In the following, we present a condensed version of the landscaping results (see Appendix B.4 for more details).

Results and Discussion. The main outcome of the landscaping analysis that contrasted the rAVS-CoO model with the AVS-BB model on our asymmetrical ROs (see Figure B.15 on page 29) is that the two models do not fully mimic each other on our asymmetrical ROs but that they still are able to closely fit the not self-generated data. However, the AVS-BB model shows a higher degree of model mimicry compared to the rAVS-CoO. That is, the AVS-BB model fits the data generated by the rAVS-CoO model almost as well as the rAVS-CoO model itself while the rAVS-CoO model shows a worse (but still good) performance for the data generated by the AVS-BB model.

Notice the general lower magnitude of the fits of the rAVS-CoO and the AVS-BB model compared to landscaping analyses that contrasted the rAVS with the rAVS-CoO model (compare Figures B.15 and B.16 with Figures B.17 and B.18 in the Appendix; pages 29f.). This reflects a higher degree of model mimicry for the AVS-BB and the rAVS-CoO model compared to the rAVS and the rAVS-CoO model. This is particular interesting because – in terms of assumptions and mechanisms about the attentional shift – the rAVS and the rAVS-CoO model are closer to each other than the AVS-BB model is to the rAVS-CoO model. Nevertheless, the AVS-BB model and the rAVS-CoO model are acting more similar to each other than the rAVS and the rAVS-CoO model. Apparently, *what* is implemented in the models (center-of-object orientation for AVS-BB and rAVS-CoO) is more important than *how* it is implemented (AVS-BB has a population of vectors, rAVS-CoO's vector sum consists of only one vector pointing in the opposite direction – in contrast, the rAVS and the rAVS-CoO model only differ in their definition of the central point).

4.5. *Discussion of Model Simulations*

Several model simulations contrasted the different implementations of attentional shifts. We fitted the models to our empirical data including a control for over-fitting (GOF, SHO). Here, the AVS model performed better than the rAVS model and the two models that favor the center-of-object over the center-of-mass (AVS-BB and rAVS-CoO) performed even better than their respective original models. The low flexibilities of the modified models (MFA) increase trust in their overall goodness: Despite the lower flexibility of the AVS-BB and the rAVS-CoO model, they provide tighter fits to the empirical data than either the AVS and the rAVS model. Most probably, this is because the two newly proposed models generate data patterns closer to the empirically observed pattern (as the PSP analysis revealed) – in particular for the asymmetrical ROs.

Using the landscaping method, we assessed whether the AVS-BB model mimics the rAVS-CoO model (or vice versa) on the asymmetrical ROs. Although both models fit the data generated by the other model, they did not fully mimic each other. Thus, in principle, our experiment was sensitive enough to distinguish between the two implemented directionalities of attention. However, since we could not do this (similar GOF and SHO performance on the asymmetrical ROs), we think that either (i) the empirical data are reliably different from all data generated by the models or (ii) the models are mimicking each other too well in the region of the empirical data. While the latter point could be addressed by designing a new empirical study or by using different empirical data (e.g., eye movements or reaction times), the first point casts doubt on the overall appropriateness of the models.

This should not be understood in the sense that the models do not fit the data well in general. The GOF results showed that all models can closely account for the empirical data. Further evidence comes from the fact that the best model fits on our whole data set have high correlations with the empirical data ($R^2 > 0.89$ for all models). Rather, the landscaping analyses point to small but reliable systematic variations in the empirical data that no model yet accounts for. The reason for this might be not properly capturing the influence of relative distance that affects central mechanisms of all current model implementations (which is why a detailed model refinement was beyond the scope of this work).

5. General Discussion

We investigated the role of shifts of (visual) attention for processing spatial language by testing predictions from models implementing contrasting directionalities of attention. Traditionally, the directionality of attention is assumed to go from the RO to the LO (Carlson, 2003; Carlson & Van Deman, 2004; Logan, 1995; Logan & Sadler, 1996; Logan & Zbrodoff, 1999). This claim, however, conflicts with recent empirical and theoretical work (Burigo & Knoeferle, 2015; Franconeri et al., 2012; Roth & Franconeri, 2012) suggesting that attention could also move from the LO to the RO. Kluth, Burigo, and Knoeferle (2017) integrated this reversed shift in a modified version of the AVS model (the reversed AVS, rAVS, model, a cognitive model rooted in the tradition of spatial language research, Regier & Carlson, 2001). However, in terms of model performance the rAVS model could not be distinguished from the AVS model using existing data. In the current article, we assessed the model performance against new human data and modified the models to capture geometric object properties that appear to matter in predicting the human spatial language acceptability ratings.

5.1. *Summary of Results*

We first estimated the predictions of the two models – the implications of the implemented assumptions – and conducted an empirical study to test them. This study revealed two as-yet-unknown sources of geometrical information that affect the evaluation of spatial language: relative distance and center-of-object orientation. The importance of the center-of-object orientation informed two modifications of the cognitive models: While the AVS and the rAVS model rely on the center-of-mass orientation in their computation, the new modifications (AVS-BB and rAVS-CoO) integrate the center-of-object orientation instead. These new modifications outperformed both the AVS and the rAVS model in terms of quantitative (GOF, SHO) and qualitative (PSP) model fits as well as model flexibility (MFA).

Comparing the AVS-BB model with the rAVS-CoO model on our rectangular ROs, slightly favors the AVS-BB model due to better fits and lower flexibility. However, neither the AVS-BB nor the rAVS-CoO model accounts for all qualitative patterns we observed in the empirical data. While results of the analyses confirmed with high probability rAVS-CoO's general prediction that relative distance affects the influence of both proximal and center-of-mass orientation, they also provided some evidence disconfirming its specific mechanism. Future model refinements should address the relative distance mechanism.

For the models at hand, the choice of using either the center-of-object or the center-of-mass is more important than the implemented directionality. This is reflected in substantial differences in model performances when comparing the AVS model with the AVS-BB model or the rAVS with the rAVS-CoO model. By contrast, model performances were virtually indistinguishable when comparing the AVS model with the rAVS model or the AVS-BB model with the rAVS-CoO model. The geometry of the RO is also what drove the looking behavior of our participants (see Appendix A.3 for analyses of the eye movements).

Using our results to answer the research question whether people prefer to direct their attention from the RO to the LO or from the LO to the RO, we are left in a position in which we cannot distinguish between these two possibilities. It seems that both directionalities may lead to equal acceptability scores. This means that we could not find evidence for the claim from Logan and Sadler (1996, p. 499) that "the viewer's attention should move *from* the reference object *to* the located object" (emphasis in the original) – but also no evidence for an opposite attentional movement.

Arguably, however, assuming an opposite movement of attention is a fundamental change to the models that should yield differing model performances. Why is it then that we could not distinguish the models and how could we still use our results to learn something about the role of attentional shifts for spatial language evaluation?

5.2. *Reasons for Non-Distinguishable Model Performances*

We identified at least three possible reasons for the non-difference in model performance. First, humans might deploy both directionalities of attention – dependent on the situation (see also discussion below). If so, the collected empirical data reflect both directionalities and neither of our models has the power to account for both directionalities (without substantial model modification). Accordingly, our suggestion for future research is to implement a model with both directionalities (and identify how either directionality is triggered).

Second, it could be that the implications of using different directionalities of attention cannot be teased apart by only looking at acceptability scores of spatial language (the direction of the attentional movement might not be reflected in these scores). Our landscaping analyses showed that the models generate different data but also that these differences are very small. Fruitful next steps would be to increase the level of detail with which the models are assessed on empirical data and perhaps model visual attention shifts more directly.

One, technical, proposal for achieving this would be by refining the output of the models, e.g., by assessing individual participants' data instead of aggregated data or by computing probability distributions across the rating range instead of just a single mean rating. The latter would allow to investigate the models' parameter spaces with the full toolkit available from the Bayesian framework (see Kluth & Schultheis, 2018, for first steps in this direction). Another idea for increasing the level of detail would be to relate the assumptions of the models about the allocation of attention on the pixel-level to fixations obtained during real-time language comprehension (i.e., to conduct a visual world paradigm study like Burigo & Knoeferle, 2015, but using stimuli more similar to the ones we have used in this article). Finally, Schultheis and Carlson (2018) showed that the computation of acceptability scores as implemented in AVS-like models interacts with the selection of a reference frame, a process currently not considered in AVS-like models. In Section 5.4, we sketch how to possibly integrate reference frames and AVS-like models .

Third, the use of LOs simplified as single points reduces the space of possible model predictions. This is because for the rAVS(-CoO) model, a single-point LO implies that the vector sum – representing the directionality of attention – de facto consists of a single vector only. This decreases the expressive power of rAVS(-CoO)'s vector sum: Holding the RO constant, the vector sum on geometrically more complex LOs differs more substantially from the RO's vector sum than does the vector sum on single-point LOs. Thus, using LOs with a mass should yield model predictions that are mutually exclusive. However, the AVS(-BB) model would need to be extended to process LOs with a mass.

5.3. *Does Directionality of Attention Matter?*

Our results are compatible with the account by Franconeri et al. (2012) who proposed that sequential shifts of attention are needed to process spatial relations. We showed that a weighted vector sum pointing from one object to the other – regardless of its directionality – successfully accounts for linguistic judgments of spatial relations. The weighted vector sum mechanism is closely related to weighted spatial pooling, a mechanism proposed to underlie the computation of saccadic endpoints (Cohen, Schnitzer, Gersch, Singh, & Kowler, 2007; Vishwanath & Kowler, 2004). Our modeling work is consistent with these low-level averaging mechanisms and suggests that such mechanisms also underlie high-level linguistic judgments.

However, different from Roth and Franconeri (2012) and Yuan et al. (2016), we could not find evidence for the claim that the linguistic asymmetry (i.e., distinguishing between the RO and the LO) is mirrored by the directionality of the attentional shift. At least in terms of linguistic acceptability judgments, both directionalities of attention accommodate a wide range of geometric effects. This does not necessarily mean that both shifts of attention are equally likely to happen in real-world scenarios. There is evidence that people shift their attention from the LO to the RO during real-time language processing (Burigo & Knoeferle, 2015). That directionality of shift also enhances reaction time performance (compared to shifting from the RO to the LO, Roth & Franconeri, 2012). In contrast, in related research by Gibson and colleagues, participants' attention shifted from the RO to the LO (e.g., Gibson & Kingstone, 2006; Gibson, Thompson, Davis, & Biggs, 2011, for review see Gibson & Sztybel, 2014).

Combining this evidence with our results (directionality seems to be irrelevant for linguistic acceptability judgments), we conclude that while an attentional shift is necessary for spatial language processing, the *directionality* of this shift seems to be flexible. Future research should clarify to what extent this flexibility is related to the task and to individual preferences.

The idea that for the processing of spatial relations the existence of asymmetries might be more important than their directionalities is consistent with studies of language development by Dessalegn and Landau (2008, 2013). In their studies, 4-year-old children had to remember visual spatial relations (the location of two colors in a square) for one second. Afterwards they had to choose the same visual relation from a set of alternatives. Compared to cueing with symmetric predicates, children's performance increased with asymmetric spatial (e.g., "the black is to the left of the white") and non-spatial terms (e.g., "the black is prettier than the white"). In particular for the non-spatial terms, the direction of the asymmetry (i.e., which color is prettier than the other) is irrelevant. Thus, marking one component of a spatial relation as different from the other component helped children to maintain their representation of the spatial relation.

5.4. Outlook: Model Refinements

For deciding in which direction attention moves, we think that other components of spatial language processing play an important role, in particular the spatial reference frame (Logan, 1995; Logan & Sadler, 1996). Spatial reference frames are a widely used representation of direction in spatial cognition in general (e.g, Pederson, 2003) and spatial language use in particular (e.g., Levinson, 2003; Majid, Bowerman, Kita, Haun, & Levinson, 2004; Schultheis & Carlson, 2017). Based on Logan and Sadler (1996), who consider reference frames to be attentional representations, Gibson and Sztybel (2014) use reference frames in their theoretical account to explain effects of linguistic cues on attentional deployment. Despite this importance for the interaction of spatial language and attention, the role of spatial reference frames in AVS-like models remains unclear. According to Logan and Sadler (1996), reference frames are three-dimensional coordinate systems that consist of four components: origin (anchor point of reference frame), scale (length of axes), orientation (rotation of axes around origin), and direction (definition of axes' end points, e.g., above vs. below). The first two components might integrate straightforwardly into AVS-like models: the models' attentional focus could be interpreted as origin and the width of the models' attentional distribution as scale.

Orientation and direction, however, require more thought. Currently, they seem to be intertwined in the models' reference direction to which the directed vector sum is compared. However, direction and orientation of both the RO (e.g., Carlson-Radvansky & Logan, 1997) and the LO (Burigo, Coventry, Cangelosi, & Lynott, 2016; Burigo & Sacchi, 2013; Burigo & Schultheis, 2018) affect spatial language acceptability judgments. Specifically, Burigo and Schultheis (2018) contrasted direction and orientation and found that the direction of the LO affects spatial language evaluation while its orientation seems to be irrelevant. This supports relating the directionality of attention in AVS-like models more to the direction component of the reference frame than to the orientation component. Future research could clarify this and investigate the interaction of spatial reference frames and directionalities of attentional shifts.

Another crucial limitation of all AVS-like models is the lack of a temporal component – in particular, since movements of attention are inherently temporal. Because no AVS-like model – regardless of the implemented directionality of attention – makes any claims about temporal aspects, we used geometrical test cases to contrast the models. Integrating a temporal component would further contrast the implications of modeling different directionalities of attention. In addition, it would better connect the models to existing temporal data like overt attentional shifts during real-time language comprehension (e.g., Burigo & Knoeferle, 2015; Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002) or reaction time data (by predicting reaction times with enhanced models; e.g., Gibson & Sztybel, 2014; Logan, 1994; Roth & Franconeri, 2012).

5.5. Conclusion

In summary, we could not settle the debate on whether people move their attention from the RO to the LO or from the LO to the RO when evaluating a spatial sentence. Implementing either directionality of attention in computational models generated predictions that we tested empirically. Based on the empirical results and further model simulations, both directionalities of attention were equally successful in accommodating the data. However, testing the predictions we found two new sources of geometric information that affect acceptability scores of spatial prepositions: relative distance and center-of-object orientation. Implementing the center-of-object orientation instead of the center-of-mass orientation, we proposed two modifications to the computational models that substantially improved their performance.

Acknowledgments

We thank Steven Franconeri for his insightful comments on this article. We would like to thank the creators and maintainers of the following open source software for making this work possible: `CGAL` (The CGAL Project, 2015), `GNU Scientific Library` (Galassi et al., 2009), `GNU octave` (Eaton, Bateman, Huauberg, & Wehbring, 2015), `R` (R Core Team, 2016), `brms` (`R` package, Bürkner, 2017), `rstan` (`R` package, Stan Development Team, 2016), `loo`

(R package, Vehtari, Gelman, & Gabry, 2016), `bayesplot` (R package, Gabry, 2017), `ggplot2` (R package, Wickham, 2009), and `gnuplot` (Williams, Kelley, & many others, 2016).

Funding. This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). The funding source was not involved in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication. The authors declare no conflicts of interest.

Author Contributions

TK, MB, and PK designed the empirical study. TK conducted the empirical study and ran the statistical analyses (consulted by MB). TK and HS conceived the model simulations. TK implemented the cognitive models and ran the model simulations (consulted by HS). TK, MB, HS, and PK wrote the paper.

Supplementary Material

The data publication [dataset]Kluth (2018, `https://doi.org/10.4119/unibi/2918231`) contains (i) all raw empirical data, (ii) R scripts to reproduce all statistical analyses and graphics, and (iii) C++ source code implementing all discussed cognitive models and model evaluation techniques. All data are licensed under the Open Database License (version 1.0, ODbL v1.0) and all source code is licensed under the GNU General Public License (version 3, GNU GPLv3).

Appendix A. Analyses of the Empirical Data

Appendix A.1. Acceptability Ratings

Table A.1 and A.2 contain the mean *über* (*above*, R1–R5) and *unter* (*below*, R6–R10) ratings of all stimuli (see Figs. 4, 5, 6, and 8 on pages 9, 10, 11, and 12 for visualizations of individual ratings, RO shapes, and row and column codings). Note that computing mean ratings assumes that the raw ratings can be interpreted on a metric scale which is strictly speaking an incorrect assumption (Liddell & Kruschke, 2018). Given that the cognitive models are fitted to mean ratings, we nevertheless provide the numbers here (but see Kluth & Schultheis, 2018, for a model extension).

Appendix A.1.1. Relative Distance: Regression Analysis

Regier (1996) and Regier and Carlson (2001) showed an effect of the center-of-mass orientation: The higher the deviation of the center-of-mass orientation (the imaginary line connecting the LO with the center-of-mass of the RO) from a reference direction, the lower the rating. Given this established effect, we expected different ratings for the four rectangles – even without considering the relative distance. This is because the taller the rectangle, the more the center-of-mass of the rectangles is moving downwards. Due to this, the

Table A.1. Mean ratings for rectangular ROs (R1–R5: *über* [*above*] ratings; R6–R10: *unter* [*below*] ratings). Cells with dashes were occupied by the RO. Note that this type of data summary assumes that the raw ratings can be interpreted on a metric scale which is strictly speaking an incorrect assumption. Given that the cognitive models are fitted to mean ratings, we nevertheless provide the numbers here (but see Kluth & Schultheis, 2018, for a model extension).

	C1	C2	C3	C4	C5	C6	C7	C8
	thin rectangle							
R1	5.12	7.50	7.79	8.38	8.56	7.97	7.59	5.59
R2	4.59	7.56	7.76	8.21	8.29	7.94	7.68	4.74
R3	3.94	7.50	7.79	8.24	8.50	7.79	7.62	3.79
R4	2.71	—	—	—	—	—	—	2.88
R5	1.47	—	—	—	—	—	—	1.38
R6	1.35	—	—	—	—	—	—	1.74
R7	2.38	—	—	—	—	—	—	2.12
R8	3.85	7.35	7.71	8.38	8.26	7.76	7.26	3.74
R9	4.62	7.44	7.74	8.21	8.18	7.88	7.47	4.59
R10	5.26	7.18	7.85	8.65	8.24	7.91	7.47	4.79
	thick rectangle							
R1	5.29	7.50	8.03	8.50	8.59	8.15	7.47	5.47
R2	4.91	7.41	7.79	8.35	8.29	8.15	7.56	5.09
R3	3.71	7.62	7.88	8.32	8.38	8.06	7.68	3.76
R4	2.97	—	—	—	—	—	—	2.62
R5	1.38	—	—	—	—	—	—	1.59
R6	1.41	—	—	—	—	—	—	1.41
R7	2.06	—	—	—	—	—	—	2.41
R8	3.62	7.32	7.88	8.03	8.21	7.65	7.53	3.41
R9	4.32	7.71	7.94	8.24	8.26	8.00	7.26	4.56
R10	4.82	7.32	7.97	8.41	8.35	7.74	7.56	5.12
	square rectangle							
R1	5.82	7.50	7.82	8.41	8.38	8.15	7.44	5.47
R2	5.00	7.50	8.12	8.09	8.26	7.82	7.56	5.06
R3	3.97	7.53	7.91	8.03	8.24	7.79	7.56	4.00
R4	2.59	—	—	—	—	—	—	2.47
R5	1.44	—	—	—	—	—	—	1.53
R6	1.56	—	—	—	—	—	—	1.38
R7	2.47	—	—	—	—	—	—	2.32
R8	3.62	7.38	7.68	8.21	8.44	7.85	7.24	3.74
R9	4.62	7.53	7.71	8.09	8.38	8.06	7.62	4.44
R10	5.12	7.65	7.76	8.44	8.29	7.82	7.74	4.94
	tall rectangle							
R1	5.62	7.82	8.15	8.50	8.41	8.12	7.41	5.35
R2	4.91	7.38	7.91	8.50	8.44	7.97	7.44	5.03
R3	4.21	7.53	7.91	8.24	8.32	7.91	7.44	3.85
R4	2.65	—	—	—	—	—	—	2.91
R5	1.41	—	—	—	—	—	—	1.71
R6	1.53	—	—	—	—	—	—	1.53
R7	2.26	—	—	—	—	—	—	2.24
R8	3.62	7.41	7.82	8.21	8.38	7.62	7.41	3.38
R9	4.59	7.41	7.82	8.32	8.41	7.91	7.56	4.79
R10	5.38	7.62	7.82	8.59	8.47	7.88	7.59	4.94

Table A.2. Mean ratings for asymmetrical ROs (R1–R5: *über* [*above*] ratings; R6–R10: *unter* [*below*] ratings). Cells with dashes were occupied by the RO or its bounding box. Note that this type of data summary assumes that the raw ratings can be interpreted on a metric scale which is strictly speaking an incorrect assumption. Given that the cognitive models are fitted to mean ratings, we nevertheless provide the numbers here (but see Kluth & Schultheis, 2018, for a model extension).

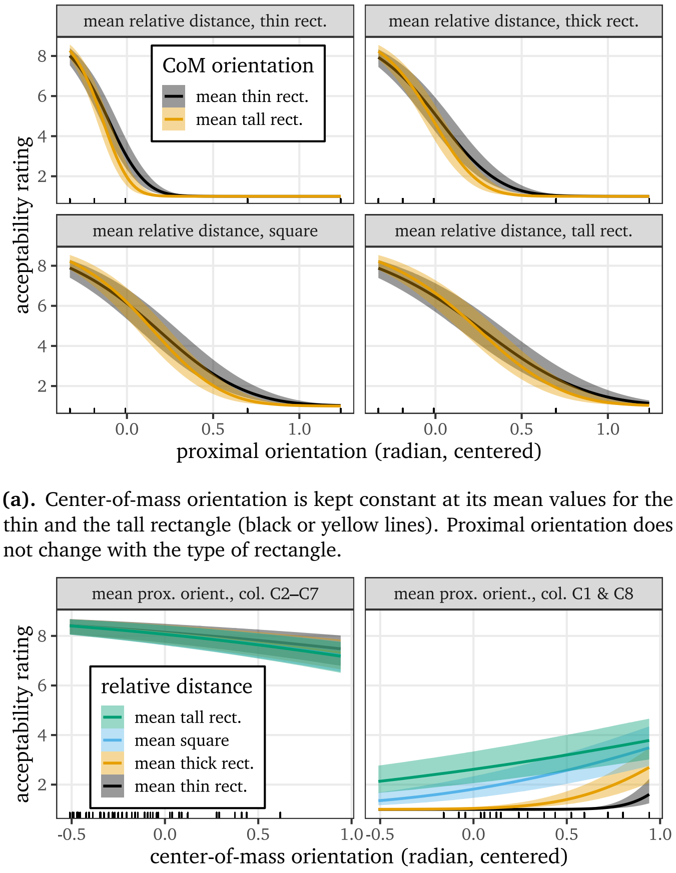
	C1	C2	C3	C4	C5	C6	C7	C8
		C RO						
R1	5.71	7.68	7.91	8.41	8.65	8.18	7.56	5.47
R2	4.88	7.76	7.85	8.47	8.24	7.94	7.53	5.12
R3	3.97	7.62	7.91	8.21	8.12	7.68	7.50	3.79
R4	2.82	—	—	—	—	—	—	2.38
R5	1.41	—	—	—	—	—	—	1.41
R6	1.53	—	—	—	—	—	—	1.50
R7	2.41	—	—	—	—	—	—	2.32
R8	4.06	7.50	7.76	8.35	8.18	7.56	7.47	4.06
R9	5.06	7.68	7.68	8.29	8.41	7.79	7.50	5.03
R10	5.06	7.56	8.03	8.56	8.56	7.94	7.53	5.38
		mC RO						
R1	5.35	7.50	7.82	8.32	8.62	8.09	7.71	5.29
R2	5.00	7.56	7.82	8.21	8.50	8.00	7.71	5.26
R3	3.82	7.65	7.50	8.00	8.47	7.79	7.56	4.00
R4	2.91	—	—	—	—	—	—	2.59
R5	1.50	—	—	—	—	—	—	1.62
R6	1.53	—	—	—	—	—	—	1.38
R7	2.24	—	—	—	—	—	—	2.53
R8	3.65	7.50	7.56	7.91	8.41	7.94	7.71	3.74
R9	5.00	7.26	7.71	8.06	8.41	7.91	7.62	5.12
R10	4.94	7.44	7.62	8.38	8.53	8.12	7.74	5.21
		L RO						
R1	5.44	7.62	8.18	8.06	8.00	8.06	7.41	5.26
R2	4.76	7.38	7.88	8.18	8.18	7.79	7.82	4.59
R3	3.59	7.59	7.91	8.24	8.21	7.91	7.53	4.00
R4	2.44	—	—	—	—	—	—	2.41
R5	1.41	—	—	—	—	—	—	1.50
R6	1.71	—	—	—	—	—	—	1.18
R7	2.29	—	—	—	—	—	—	2.53
R8	3.79	7.76	8.03	8.38	8.29	7.76	7.29	3.35
R9	4.06	7.44	8.00	8.15	8.24	7.76	7.53	4.68
R10	5.06	7.47	7.97	8.29	8.38	8.09	7.50	4.97
		mL RO						
R1	5.38	7.41	7.68	8.18	8.47	7.91	7.62	5.21
R2	4.79	7.56	8.00	8.03	8.38	7.74	7.56	4.94
R3	4.24	7.26	7.56	8.03	8.35	7.59	7.50	3.65
R4	2.71	—	—	—	—	—	—	2.50
R5	1.47	—	—	—	—	—	—	1.62
R6	1.44	—	—	—	—	—	—	1.50
R7	2.21	—	—	—	—	—	—	2.24
R8	3.71	7.44	7.68	8.12	8.35	7.94	7.68	3.62
R9	4.68	7.35	7.59	8.12	8.32	7.74	7.74	4.47
R10	4.85	7.32	7.68	8.53	8.47	8.12	7.79	5.21

center-of-mass orientations of the LOs above the rectangles reduce with the height of the RO. Since a lower center-of-mass orientation should lead to a higher rating, we would expect to find higher ratings for the taller rectangles compared to the shorter rectangles. However, we did not find such an effect (see Figure 3) suggesting that the relative distance influences the extent to which the center-of-mass orientation affects acceptability ratings.

To investigate whether the assumption about the role of the relative distance from the rAVS model are reflected in our empirical data, we used the metric predictor relative distance in an ordinal regression to predict the ratings. We used equation 1 from the rAVS model to compute the relative distance as the AVS model provides no explicit definition of relative distance. Furthermore, the regression model includes the predictors center-of-mass orientation and proximal orientation (in radian notation to facilitate model convergence) and allowed full interactions between all three predictors. We compared all possible alternative models using these predictors (e.g., not allowing interactions or removing certain predictors) with the LOO method to ensure that this most complex model is doing considerably better on the data than simpler models without over-fitting. Furthermore, we fitted the same most complex model to a data subset consisting only of LOs on the "correct" side of the grazing line (i.e., excluding rows R4–R7). While the different data sets resulted in different estimated regression slopes, the qualitative results of both models remained equal. Therefore, we present the model fitted to our whole data set for the rectangular ROs. We centered all predictors around their mean.

The results of the full interaction model can be best understood by depicting the effects of all three predictors on the outcome variable as estimated by the regression model. Figure A.10 plots such visualizations of the same regression model from different perspectives. Since all predictors are metric, we needed to keep some predictors constant at discrete levels in order to plot the estimations of the regression model. In Figure A.10a, we kept the predictor center-of-mass orientation constant on two levels (cf. colored lines): The mean center-of-mass orientation for all LOs above/below the thin (comparably large center-of-mass orientation) and the tall rectangle (comparably small center-of-mass orientation). In Figure A.10b, the proximal orientation is held constant on two levels (cf. subplots in Figure A.10b): the mean proximal orientation for LOs directly above/below the RO (where the proximal orientation does not deviate from the reference direction; columns C2–C7, see Figures 4 and 5) and the mean proximal orientation of LOs in columns C1 and C8 (where the proximal orientation does deviate from the reference direction). For all plots, we kept the value of the predictor relative distance constant on four levels (cf. subplots in Figure A.10a and colored lines in Figure A.10b): the mean relative distance for all LOs above/below one of the four rectangles (with lower relative distance for taller rectangles). All y-axes denote the predicted variable acceptability rating.[12]

[12] A caveat: In the plots, ratings are considered to be continuous, which is



(a). Center-of-mass orientation is kept constant at its mean values for the thin and the tall rectangle (black or yellow lines). Proximal orientation does not change with the type of rectangle.

(b). Proximal orientation is kept constant at two levels: mean proximal orientation for LOs in columns C2–C7 (left subplot) and mean for LOs in columns C1 & C8 (right subplot).

Fig. A.10. Effects (and interactions) of relative distance, center-of-mass orientation, and proximal orientation on rating as estimated by an ordinal regression model using the data from the rectangular ROs. Panel (a) keeps the center-of-mass orientation constant at two levels, panel (b) keeps the proximal orientation constant at two levels. For both panels, relative distance is held constant at four levels (its mean value for each of the four rectangular ROs). Black bars at the bottom of each plot depict the values of the corresponding predictor tested in the experiment (in the given condition). Shaded areas denote 95% credible intervals of the estimate. For convenience, these plots assume the outcome (rating) to be metric which is not how they are treated by the regression model. Hence, these plots should only be used to intuitively grasp the effects and interactions of the different predictors on the outcome.

Our main interest was in whether the relative distance affected the strength of the center-of-mass or proximal orientation. Across all four subplots in Figure A.10a, higher proximal orientation correlates with lower ratings (cf. negative slopes). However, this (expected) effect of proximal orientation reduces with decreasing relative distance as revealed by steeper slopes for high relative distance (thinner rectangles; e.g., top left subplot) compared to low relative distance (taller rectangles; e.g., bottom right subplot). This is to say that the smaller the relative distance, the lower the influence of proximal orientation on ratings.

In all subplots of Figure A.10a, the two colored lines of center-of-mass orientation cross each other due to their different slopes (the yellow line is steeper than the black line). This means that for small center-of-mass orientations (tall rectangle, yellow line), the proximal orientation has a greater influence than it has for large center-of-mass orientations (thin rectangle, black line). This effect on the influence of proximal orientation is less pronounced than the effect of relative distance while it goes in the opposite direction: While lower relative distance (i.e., taller rectangles) co-occurs with lower influence of proximal orientation (less steep slopes compared to higher relative distance, cf. subplots in Figure A.10a), lower center-of-mass orientations (again for taller rectangles) strengthen the impact of proximal orientation (steeper slopes than for higher center-of-mass orientations, cf. colored lines in Figure A.10a).

Figure A.10b keeps the proximal orientation constant and allows to analyze the effect of center-of-mass orientation on acceptability rating more closely. For LOs directly above/below the rectangles (columns C2–C7, see left subplot of Figure A.10b), the proximal orientation does not deviate from the reference direction. For these LOs, participants generally gave high ratings and the center-of-mass orientation shows the expected effect: Higher center-of-mass orientation correlates with lower ratings (negative slope). Different relative distances do not change this influence of the center-of-mass orientation (all four colored lines overlap). For LOs placed in columns C1 or C8, this picture changes drastically (right subplot of Figure A.10b). For these LOs, the proximal orientation does deviate from the reference direction and overall our participants rated these LOs considerably lower than LOs in columns C2–C7. More interestingly, however, is that for these LOs the effect of the center-of-mass orientation is reversed: The larger the center-of-mass orientation, the higher the rating (positive slopes in the right subplot of Figure A.10b). Moreover, relative distance affects the reversed influence of the center-of-mass orientation: The larger the relative distance (i.e., the thinner the rectangle), the lower are the ratings for high center-of-mass orientations (in the right subplot of Figure A.10b: steeper slopes and lower values for large relative distances, i.e., thinner rectangles, compared to small relative distances, i.e., taller rectangles).

Appendix A.2. Asymmetrical ROs: Center-of-Object Orientation

As another test of our suggestion that people rely on the center-of-object orientation we compared ratings for LOs that share the same center-of-object orientation on average. For this comparison, we split the ratings for the asymmetrical ROs in two subsets: ratings for LOs to the left of the center-of-object (columns C1–C4) and ratings for LOs to the right of the center-of-object (columns C5–C8). Then, we predicted the rating based on the location of the LO (left or right). If people use the center-of-object orientation, we should find equal ratings for these subsets. Using the default prior for the slope coefficient, the posterior distribution of the model

not a valid assumption for the underlying ordinal regression. Nevertheless, the plots illustrate the impact of the predictors on the magnitude of the rating.

confirms that there is no difference in ratings for LOs on the left or on the right side ($\beta_{right} = 0.05$, 95% CI $[-0.03, 0.13]$).

However, since we balanced the side of the cavity in our asymmetrical ROs to control for a possible left-right bias, the previous analysis collapsed across the influence of the location of the “mass” side of the RO. So, our next analysis predicted rating based on whether the LO was on the side of the center-of-mass (i.e., columns C1–C4 for ROs L and C and columns C5–C8 for ROs mC and mL). Here, we found a credible but small influence of the distribution of mass. LOs on the same side as the center-of-mass received higher ratings than LOs on the opposite side ($\beta_{CoMSide} = 0.10$, 95% CI $[0.02, 0.18]$, default priors). More precisely, the regression model assigns a 2% higher probability for ratings 8 and 9 if the LOs were on the side of the center-of-mass than for LOs on the other side. This is to say that the distribution of mass has an effect on the ratings but it is certainly not as high as would be expected if people rely only on the center-of-mass orientation.

Appendix A.2.1. Replication: Über Versus Unter

Researchers reported lower acceptability ratings for inferior prepositions (*below, under*) compared to superior prepositions (*above, over*; e.g., Burigo & Coventry, 2005; Burigo et al., 2016; Carlson & Logan, 2001). These studies presented data from English native speakers, whereas our study was conducted with German native speakers. We were interested to what extent the reported effect generalizes to the corresponding German prepositions *über* and *unter*.

Burigo et al. (2016) used acceptability rating studies with the same rating scale as in our study. Their main interest concerned other aspects of spatial language processing but they also found significantly lower ratings for the prepositions *below/under* compared to ratings for *above/over*. We used their data to estimate $\mu = -0.11$ for the prior distribution of the slope parameter in the regression on our data. Since we used a Gaussian distribution as prior distribution, we also needed to specify a standard deviation. The studies that informed our prior were conducted in English but our study was done in German, so we specified a rather large standard deviation of 0.2. This gives considerable amount of probability for values of the slope being 0.0 or even positive (denoting a reversed effect) which reflects our uncertainty whether the effect found for English is also present for German.

Running this model, we obtained a 95% credible interval of the posterior distribution for the *preposition* slope ranging from -0.115 to -0.001 with a mean of -0.058. That is, the slope is with 95% probability below zero, which means that participants in our study gave lower ratings for *unter* than for *über*. The regression model can be used to further quantify this statement by reporting the probabilities of each rating dependent on the preposition: The rating 9 was chosen with 1% more probability if the preposition was *über* compared to *unter*. This is a very small effect (as could be already seen in the small magnitude of the slope) but roughly in the same magnitude as that reported in previous literature. Taking a more conservative standpoint, we ran the same model again with the default prior as implemented in the `brms` package. This default prior is meant to be non-informative. The 95% credible interval of this model ranges from $[-0.1162, -0.0002]$ with a mean of -0.0574. Note that the upper boundary of the credible interval is closer to zero compared to the first model due to the non-informative prior. Still, the interval does not contain 0.0, suggesting lower ratings for *unter* than for *über*. Again, the probability of rating 9 is 1% higher for *über* compared to *unter*. Accordingly, our study generalizes the effect that superior prepositions are rated higher than inferior prepositions from English prepositions to German prepositions.

Appendix A.2.2. Replication: Grazing Line Effect

Regier and Carlson (2001) reported that LOs above the grazing line are rated significantly higher than points below or on the grazing line (for *above*; vice versa for *below*). The grazing line is the imaginary line that touches the top (or bottom for *below*) of the RO. For *über* (*above*), we compared the ratings for the four LOs below or on the grazing line (rows R4 and R5) with the ratings for the four LOs above the grazing line (rows R2 and R3, columns C1 and C8) across all ROs (see Figures 4, 5, 6, and 8). For *unter* (*below*), we compared the ratings for the four LOs above or on the grazing line (rows R6 and R7) with the ratings for the four LOs below the grazing line (rows R8 and R9, columns C1 and C8) across all ROs. We predicted rating as a function of being on the corresponding side of the grazing line (i.e., above the line for *über*, *above*; below the line for *unter*, *below*) or on the non-corresponding side (below or on the line for *über*, *above*; above or on the line for *unter*, *below*). Based on the summary statistics of the grazing line effects in exp. 5 and 6 reported in Regier and Carlson (2001), we specified a Gaussian distribution with $\mu = 3.7$ and $\sigma = 3.0$ as an informed prior distribution for the slope parameter of the model. This prior distribution reflects our knowledge about the grazing line effect while having a fairly high σ value due to the methodological difficulties of using mean differences of ratings as a prior for an ordinal regression. The posterior distribution of the slope, however, is quite peaky and narrow around its mean of 3.49 with a 95% credible interval ranging from 3.34 to 3.65. For LOs on the non-corresponding side of the grazing line (rows R4–R7), rating 1 has a 41% higher probability than it has for LOs on the corresponding side of the grazing line. Clearly, people more often used rating 1 for LOs on the non-corresponding side compared to LOs on the corresponding side. Running the same model using the default non-informative priors only slightly changes the estimates but not the general qualitative outcome. This replicates the effect reported by Regier and Carlson (2001).

Appendix A.3. Eye Movements

Eye movement data are in particular interesting, as they add another empirical measure (apart from ratings) that could be used to benchmark the AVS and the rAVS model. Eyes were only tracked during the picture display but not during the display of the sentence. Since we recorded eye movements

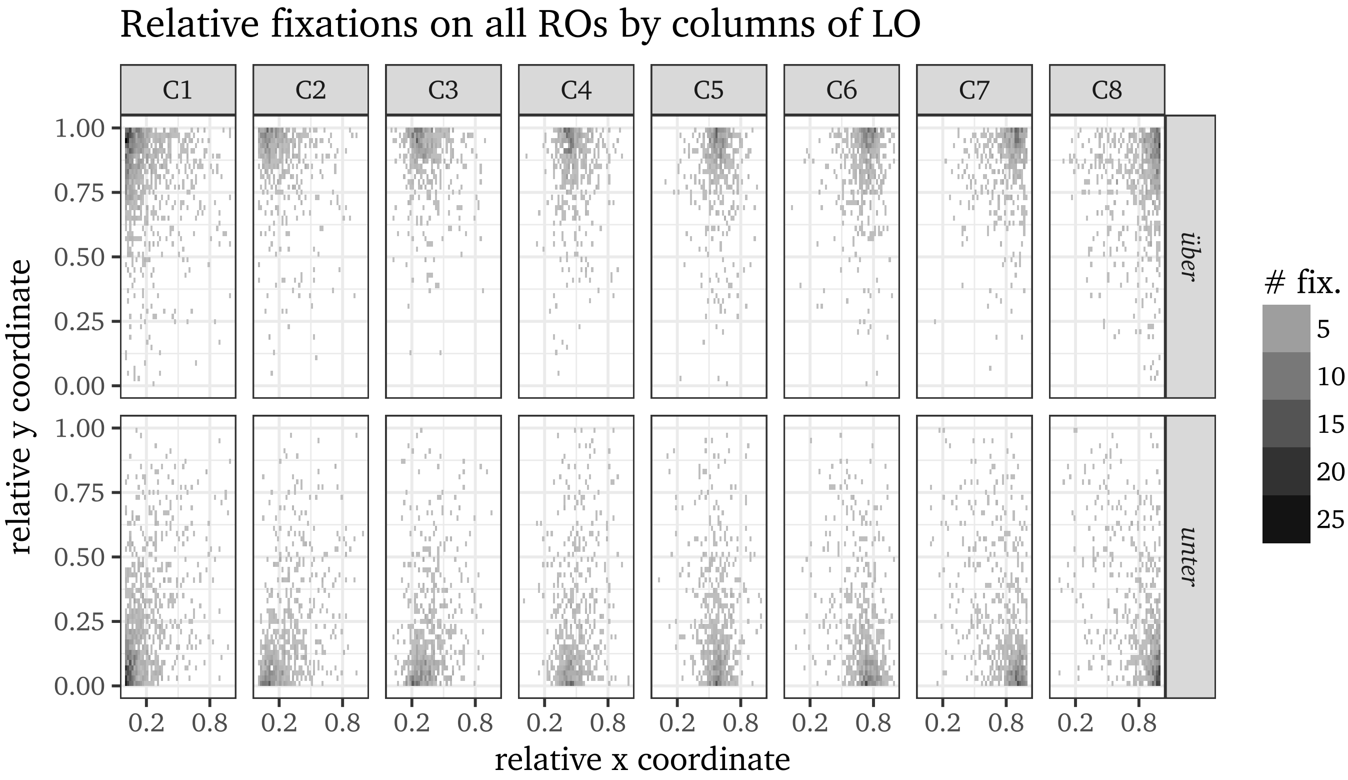


Fig. A.11. Heatmaps of relative fixations in the bounding boxes of all ROs grouped by column of the presented LO and preposition. A relative location of 0.0 corresponds to the left side (x coordinate) or the bottom (y coordinate) of each RO, a 1.0 to the right side (x coordinate) or the top (y coordinate). Darker color means more fixations (see legend). We included only fixations starting later than 150 ms after display onset.

after the participants read the sentence, we cannot directly link the fixations to the processing of the spatial prepositions in terms of movements of attention during spatial language understanding. However, we can investigate how people inspect a scene in order to verify whether it matches the spatial utterance. More precisely, we were interested in whether people fixate one specific point more than other points of the RO: the attentional focus as defined in the AVS model. The rAVS model also bases its computation on this point. We are only aware of Carlson et al. (2006) who also explicitly link assumptions about attentional allocation in AVS-like models to overt attentional behavior. Our second analysis contrasted fixations to the center-of-mass or center-of-object of the asymmetrical ROs.

Data Set. Since many of the fixations before 150 ms landed close to the center-of-mass of the RO (a region we were especially interested in) and the planning of a saccade takes approximately 200 ms (Matin, Shao, & Boff, 1993, cited in Tanenhaus, Spivey Knowlton, Eberhard, & Sedivy, 1995), we analyzed only fixations that started 150 ms after the presentation of the RO.

This leaves us with 53 718 fixations (per subject M = 1 579.94, SD = 688.76; per trial and subject M = 3.53, SD = 2.82). Out of these 53 718 fixations, ca. 46% landed close to the current LO (i.e., within a ca. 1.82 degrees of visual angle or 90 pixel wide square centered at the LO) and ca. 21% landed inside the bounding box of the RO.[13] Since our LOs were rather small we cannot investigate any differences in fixations inside the LO. Our ROs, however, were considerably larger, allowing us to explore the looking behavior over their surface. In the following, we focus on the fixations inside the bounding boxes of the ROs.

Attentional Focus from the AVS Model. The AVS model defines the attentional focus to be the point at the top of the RO that is vertically aligned with the LO (for *above*; for *below*, the focus lies on the corresponding point at the bottom of the RO). Due to our experimental design, however, we cannot analyze the vertical component of the assumed attentional focus. This is because we placed no LO below the RO for *über* (*above*; and no LO above the RO for *unter*, *below*). The worst example for every preposition was just below (or above) the top (or bottom) of the RO (rows R5–R8). These LOs, however, were still quite close to the top (or bottom) of the RO. Although we found a strong preference to look at the top of the RO for *über* (*above*; and at the bottom for *unter*, *below*), we cannot reliably tell if this difference in fixation locations actually originates from the used preposition or from the location of the LO.

[13] Note that the LOs in the rows R3–R8 were only ca. 0.3 degrees of visual angle (15 pixels) away from the RO and the precision of the eye tracker is of similar magnitude. Thus, some fixations for these trials were counted twice for the proportions: close to the LO and inside the bounding box of an RO.

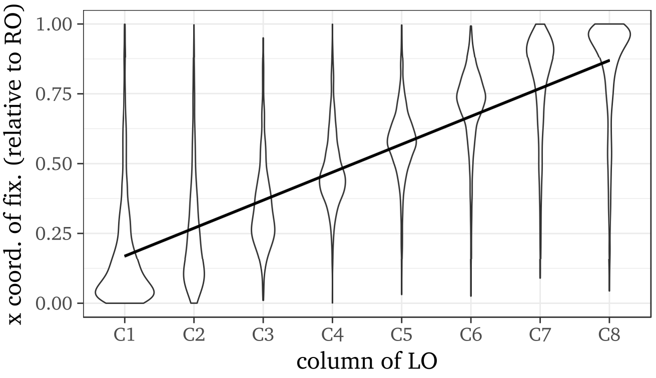


Fig. A.12. Horizontal component of relative locations of fixations in the bounding boxes of the ROs plotted by column of the LO. A relative location of 0.0 corresponds to the left side of the RO, a 1.0 to the right side. The wider the distribution, the more fixations. The line is a regression lines for predicting the relative location of fixation based on the horizontal location of the LO ($R^2 = 0.67$). We included only fixations starting later than 150 ms after display onset.

Accordingly, we just present an analysis of the horizontal component of the assumed attentional focus: a point on the RO that is vertically aligned with the location of the LO. In Figure A.11 we plotted heatmaps of all fixations inside the bounding boxes of all ROs – grouped by preposition and by the column of the LO. Fixations are coded relative to the bounding box of the RO: 0.0 means the left (or bottom) of the bounding box, 1.0 means the right (or top). The heatmaps show that the horizontal location of the fixations corresponds to the horizontal location of the LO. Figure A.12 shows only the horizontal component of all fixations grouped by the column of the LO that was shown during that trial. The wider the distribution for each group in Figure A.12, the more fixations were counted at that location. To investigate whether our participants fixated close to a vertically aligned point, we ran a linear regression asking whether the horizontal location of the LO (i.e., the column of the LO) predicts the relative x-coordinate of the fixations in the bounding box of the RO. We plotted the regression line with an $R^2 = 0.67$ on top of the data in Figure A.12. It can be seen that the horizontal location of the LO is a good predictor of the horizontal location of the fixations in the RO. This means that if the LO was shown on the left side of the screen, our participants fixated the left side of the RO more often than the right side. If the LO was shown on the right side of the screen, however, they fixated more on the right side of the RO (and respectively for the middle of the screen).

Using the x-coordinate of the LO as (centered and scaled) metric predictor instead of the categorical predictor LO-column, we specified a Bayesian regression model to predict the horizontal component of the relative fixation. The posterior distribution of this regression model supports our previous observation: The x-coordinate of the LO did reliably affect the horizontal fixation location ($\beta_{LO_x} = 0.241$, 95% CI $[0.237, 0.245]$). To exclude the possibility that fixations to LOs close to the RO were falsely classified as fixations on the RO, we ran a second regression model that excluded fixations from trials with such LOs (i.e., we excluded all trials with LOs in rows R3–R8). This regression model resulted in a slightly lower estimate for the regression coefficient while replicating the general observation ($\beta_{LO_x} = 0.203$, 95% CI $[0.196, 0.210]$).

Taken together, we found that our participants fixated locations on the RO that were vertically aligned with the LO. Such a location corresponds to the horizontal component of the attentional focus defined in the AVS that also plays a crucial role in the rAVS model. This provides empirical support for assumptions of attentional allocation made in both models.

Center-of-Mass Versus Center-of-Object. Considering the unexpected finding in the analysis of the acceptability ratings suggesting that our participants relied more on the center-of-object orientation instead of on the center-of-mass orientation, we were interested whether we can see a similar preference for any of the two centers in the fixation data.

To contrast the number of fixations to either the center-of-mass or the center-of-object for the asymmetrical ROs, we defined ca. 1.01 degrees of visual angle (50 pixel) wide squared areas of interest, one around each of the two centers. That is, we counted the number of fixations that were not farther away than ca. 0.51 degrees of visual angle (25 pixels) in either direction from either center. The number of fixations that landed close to one of the two centers can be found in Table A.3, collapsed over the used preposition.

First of all, we note that not many fixations landed in the two areas. Out of 6 193 fixations inside the bounding box of the asymmetrical ROs in total, only 200 ($\sim$ 3%) were close to one of the two centers. For the two L-shaped ROs there is a clear bias towards the center-of-mass (more than 85% of the fixations close to any of the two centers landed near the center-of-mass). This bias can be explained by the fact that for the L-shaped ROs the center-of-mass lies closer to the top or bottom of the RO than the center-of-object (see Figure 8, page 12). We already established that people fixate the top or bottom of the RO more than other parts of the RO, so it is no surprise that people also fixate a region closer to this attractor more.

The locations of the two centers for the C-shaped ROs, however, share the same y-coordinate. This means that both centers have the same distance to both the top and the bottom of the RO and the number of fixations close to the centers should not be influenced by the fact that people mostly fixate the top or bottom of the RO. Indeed, the number of fixations at the C-shaped ROs show a different picture (see Table A.3). Despite the fact that the center-of-mass is inside the RO and the center-of-object is outside the RO (but inside the bounding box of the RO), we see more fixations close to the center-of-object than to the center-of-mass.

To overcome the problem with the different vertical locations of the two centers for the L-shaped ROs and in order to draw on a greater subset of fixations, we analyzed whether our participants had a general bias to look more on the left or on the right side of the RO. To do this, we compared the num-

Table A.3. Absolute and relative number of fixations in a ca. 1.01 degrees of visual angle (50 pixel) wide square centered at the center-of-object or center-of-mass of the RO.

	C	mC	L	mL
center-of-mass	15 34.1%	27 47.4%	53 86.9%	34 89.5%
center-of-object	29 65.9%	30 52.6%	8 13.1%	4 10.5%

Table A.4. Absolute and relative number of fixations inside the bounding boxes of the ROs split by left or right landing positions.

	left		right	
thin	508	51.9%	471	48.1%
thick	618	51.7%	577	48.3%
square	699	49.8%	705	50.2%
tall	785	50.3%	777	49.7%
C	810	48.0%	879	52.0%
mC	814	48.9%	851	51.1%
L	876	60.2%	579	39.8%
mL	573	41.4 %	810	58.6%

ber of fixations that landed on the left side of the RO (relative x-coordinate smaller than 0.5) with the number of fixations that landed on the right side of the RO (relative x-coordinate greater than 0.5). We ignored fixations that were directly in the middle of the RO (relative x-coordinate of fixations equals 0.5, this affected only 3 fixations).

The upper part of Table A.4 shows the number of fixations for the rectangular ROs. Fixations to these ROs provide a baseline, as they are not asymmetric. Accordingly, we expected no bias to look at either side of the RO (apart from a potential general left-right bias in looking behavior). Indeed, our participants did not prefer either side of the rectangular ROs. The lower part of Table A.4 shows the number of fixations on either the left or the right side for the four asymmetrical ROs. Here, the data seem to provide no evidence for a left-right bias for the C-shaped ROs. Both sides of the C RO and the mC RO are fixated to an equal amount – despite the asymmetrical distribution of mass in the C-shaped ROs. In particular, this means that fixations are not biased to the location of the center-of-mass. People seem to look differently at the L-shaped ROs, though. For the L RO, the proportions of fixations suggest a preference to the left side and for the mL RO, a preference to the right side of the RO. On both preferred sides the center-of-mass of the RO is located. That is, our participants preferred to look in the direction of the center-of-mass compared to the center-of-object for the L-shaped ROs. Apparently, the more open shape of the L-shaped ROs leads to a different pattern of fixations.

To statistically support these observations, we ran a Bayesian regression model that predicted the relative x-coordinate of fixations as a function of the RO. This regression model supports our interpretation of the proportions shown in Table A.4. The predicted average horizontal fixation location for the thin rectangle (the intercept of the regression model) is not credibly different from 0.5 ($\beta_{thin} = 0.49$, 95% CI $[0.47, 0.51]$) indicating no preference for either side of this RO. None of the regression coefficients for the other rectangular ROs (coding for different fixation locations compared to the thin rectangle) is credibly different from zero meaning no preferred side of fixations for any of the rectangular ROs ($\beta_{thick} = -0.01$, 95% CI $[-0.04, 0.01]$, $\beta_{square} = 0.01$, 95% CI $[-0.02, 0.03]$, $\beta_{tall} = 0.00$, 95% CI $[-0.02, 0.03]$). More interestingly, the regression coefficients for the C and mC ROs do also not differ credibly from zero ($\beta_C = 0.01$, 95% CI $[-0.01, 0.04]$, $\beta_{mC} = 0.01$, 95% CI $[-0.01, 0.03]$). This indicates that our participants looked at these C-shaped ROs like they were rectangles, i.e., like they had no cavities. By contrast, both regression coefficients for the L-shaped ROs were credibly different from zero ($\beta_L = -0.05$, 95% CI $[-0.07, -0.02]$, $\beta_{mL} = 0.06$, 95% CI $[0.03, 0.08]$). The sign of these regression coefficients corresponds to the location of the "leg" of the L-shaped ROs. That is, for the L-shaped ROs, our participants looked more on the side of the RO with a greater amount of mass (the side where the center-of-mass is located).

Appendix A.3.1. Discussion: Eye Movements

We found that the horizontal component of participant's fixations matched the horizontal component of the assumed point F of maximal attention (as defined in the AVS model and as used in the rAVS model). This supports the assumptions about allocation of attention in both models and thus strengthens the linking hypothesis of attention as used in the model and overt attention as measured by an eye-tracker. Finally, we found different looking behaviors when comparing the C-shaped ROs with the L-shaped ROs. Our participants fixated the C-shaped ROs as if they were rectangles, while they fixated the "legs" of the L-shaped ROs more than the cavities, i.e., the asymmetry of the L-shaped ROs was reflected in the fixations. The fixation pattern for the L-shaped ROs (reflecting the asymmetrical distribution of mass) is in line with the prominent role of the center-of-mass for perceptual and saccadic localization (e.g., Desanghere & Marotta, 2015; Melcher & Kowler, 1999; Nuthmann & Henderson, 2010; Vishwanath & Kowler, 2003; note that Melcher & Kowler, 1999, found that the center-of-mass of the abstract shape was preferred even if the visible density of points making up the shapes was manipulated). In contrast, the fixation pattern for the C-shaped ROs (ignoring the asymmetrical distribution of mass) highlights the influence of the particular task on looking behavior (cf. discussion in Vishwanath & Kowler, 2003, who speculate about the role of reference frames and spatial pooling processes similar to the vector sum; see also discussion in Melcher & Kowler, 1999). Future research should more precisely identify how linguistic tasks affect preferred fixation locations.

Appendix A.4. Reaction Times

All regression models that we used to analyze reaction times used an exponentially modified Gaussian distribution as response distribution – a common choice for modeling reaction times (Dawson, 1988; Van Zandt, 2000). Note that our task was self-paced and we did not ask participants to respond as quickly as possible. Thus, we did not expect large differences in reaction times.

Replication: Über Versus Unter. Researchers have reported shorter reaction times for *above* compared to *below* (e.g., Carlson & Logan, 2001, note 1). We were interested, whether this effect generalizes to German (quicker responses for *über* compared to *unter*). To analyze this, we specified a regression model that predicted reaction time from preposition. Our participants responded slightly more quickly to *über* trials (mean = 1857.73 ms) than to *unter* trials (mean = 1873.34 ms) as revealed by the regression coefficient for preposition that is credibly different from 0 ($\beta_{unter} = 17.40$, 95% CI $[4.76, 29.61]$). These results generalize the effect that people respond faster to superior prepositions than to inferior prepositions from English to German.

Rows. A second analysis predicted reaction time as a function of the row of the LO. Compared to reaction times for LOs in row R1 (the top row; mean = 1821.72 ms), we found credibly longer reaction times for LOs in rows R4–R7 ($\beta_{R4} = 192.28$, 95% CI $[149.25, 234.71], M_{R4} = 2360.34$ ms; $\beta_{R5} = 57.42$, 95% CI $[16.65, 98.44], M_{R5} = 2100.58$ ms; $\beta_{R6} = 48.55$, 95% CI $[6.65, 89.32], M_{R6} = 1953.16$ ms; $\beta_{R7} = 158.88$, 95% CI $[115.89, 200.74], M_{R7} = 2325.54$ ms) as well as slightly shorter reaction times for row R2 ($\beta_{R2} = -24.59$, 95% CI $[-48.91, -0.06], M_{R2} = 1800.72$ ms). In particular the longer reaction times for row R4–R7 can be readily explained by the location of the LOs: All LOs in these rows are either on the grazing line (see Appendix A.2.2) or on the side of the grazing line that does not correspond with the preposition people should rate.

Columns. With the next regression model, we predicted reaction time as a function of the column of the LO. Compared to LOs in column C1 (mean = 2260.64 ms), people responded faster to the LOs in columns C2–C7 ($\beta_{C2} = -220.85$, 95% CI $[-247.56, -194.37], M_{C2} = 1762.72$ ms; $\beta_{C3} = -242.25$, 95% CI $[-269.58, -215.79], M_{C3} = 1646.78$ ms; $\beta_{C4} = -283.33$, 95% CI $[-309.96, -257.15], M_{C4} = 1559.70$ ms; $\beta_{C5} = -286.06$, 95% CI $[-313.23, -259.60], M_{C5} = 1547.57$ ms; $\beta_{C6} = -261.47$, 95% CI $[-288.38, -234.76], M_{C6} = 1653.74$ ms; $\beta_{C7} = -202.07$, 95% CI $[-228.11, -175.59], M_{C7} = 1759.41$ ms) but not to LOs in column C8 ($\beta_{C8} = -4.09$, 95% CI $[-26.52, 18.21], M_{C8} = 2228.42$ ms). That is, our participants were quicker to rate LOs with non-deviating proximal orientation (columns C2–C7) compared to LOs with a proximal orientation greater than 0 (columns C1 & C8).

Appendix B. Model Simulations

Appendix B.1. Fitting Algorithm and Parameter Ranges

The Root Mean Square Error (RMSE) is defined as follows

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(data_i - modelOutput_i)^2} \tag{B.1}$$

The upper limit of the RMSE depends on the range of the underlying data. Due to the rating scale we used in our experiment (from 1–9), the worst value of an RMSE for our data would be 8.0 (e.g., if all participants rated everything with a 9 but the model computes only 1s). Other studies used different rating scales and therefore also the worst value of the RMSE shifts. Regier and Carlson (2001) for instance used a rating scale from 0 to 9 which results in the worst possible RMSE value of 9.0. To be able to compare RMSE values throughout different rating studies we applied the normalized RMSE (nRMSE), defined as:

$$nRMSE = \frac{RMSE}{rating_{max} - rating_{min}} \tag{B.2}$$

The nRMSE always produces a value between 0.0 and 1.0 (with 0.0 being the best and 1.0 being the worst value). In order to find the lowest nRMSEs, we applied a parameter estimation technique known as simulated annealing, a variant of the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953).

Model Parameters. The AVS model has four free model parameters: λ (controls the width of the attentional distribution), $slope$ and $intercept$ (qualifying the linear mapping of angular deviation to acceptability score), and $highgain$ (modifying the angular component as a function of vertical LO-to-RO position).

The rAVS model has four free parameters of which three are the same as in the AVS model ($slope, intercept, highgain$). However, the rAVS model does not use the λ parameter (Kluth, Burigo, & Knoeferle, 2017, p. 290). Instead it qualifies the strength of the effect of the relative distance on the vector direction with its parameter α. We used the following ranges for the model parameters:

$$\frac{-1}{45} \leq slope \leq 0 \tag{B.3}$$

$$0.7 \leq intercept \leq 1.3 \tag{B.4}$$

$$0 \leq highgain \leq 10 \tag{B.5}$$

$$0.001 \leq \lambda \leq 5 \tag{B.6}$$

$$0.001 \leq \alpha \leq 5 \tag{B.7}$$

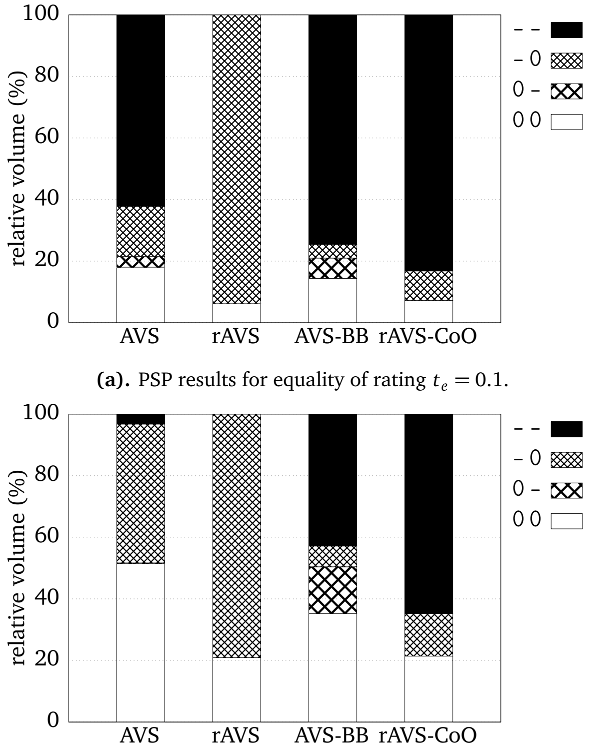
Appendix B.2. Parameter Space Partitioning (PSP)

Method. To quantify the range of qualitatively different model predictions, we simulated the models using the Parameter Space Partitioning (PSP) algorithm developed by Pitt et al. (2006). The main idea of the PSP is the following: Given the stimuli under consideration (see Figure 2), enumerate all possible model tokens (a model with a fixed set of parameters, cf. Wagenmakers et al., 2004), simulate all these model tokens and categorize the different outputs of the model into qualitatively different output patterns (i.e., model predictions). The PSP algorithm estimates the volumes that these model predictions cover in the parameter space. Internally, the PSP algorithm is a Markov Chain Monte Carlo (MCMC) approach (which makes it faster than a naïve complete enumeration of the parameter space; for more details about the algorithm see Kim, Navarro, Pitt, & Myung, 2004 and Pitt et al., 2006).

We slightly changed the MATLAB implementation made available by Pitt et al. (2006) under http://faculty.psy.ohio-state.edu/myung/personal/psp.html in order to be able to use it with GNU octave (Eaton et al., 2015). As input for the models, we used the stimuli in Figure 2. To increase the reliability of the measurement, we added more LOs above each RO (28 LOs for the rectangular ROs, 12 LOs for the asymmetrical RO) and contrasted mean ratings. If the mean rating for the thin rectangle was lower than for the tall rectangle, we coded this as “-”. If LOs above the thin rectangle were rated higher than LOs above the tall rectangle, we coded this as “+”. If LOs above both rectangles got the same ratings, we coded this as “0”. For the test case with the asymmetrical ROs, we compared the mean ratings for 6 LOs placed to the left of center-of-mass of the L-shaped RO with 6 LOs placed to the right of the center-of-mass. We used the same coding for this comparison with “-” meaning lower ratings for LOs on the left side compared to the right side, “+” meaning higher ratings for left LOs and “0” meaning equal ratings. Since we have two test cases for predictions (relative distance case and asymmetrical ROs case), we obtained a two-digit code. The first digit codes the relative distance test case and the second digit codes the test case with the asymmetrical ROs. We defined two thresholds t_e for equality of ratings (informed by significant mean differences in spatial language acceptability ratings reported by Burigo et al., 2016; Carlson-Radvansky, Covey, & Lattanzi, 1999; Hörberg, 2008; Regier & Carlson, 2001). If two mean ratings differ by less than $t_e = \{0.1, 0.5\}$ they are considered to be equal. In these simulations, the rating scale used by the models ranged from 1.0 to 9.0 (matching the rating scale applied in our empirical study).

Appendix B.2.1. PSP Results: AVS and rAVS Models

We ran the PSP algorithm three times for every threshold t_e and report the mean estimates of relative volume that every generated pattern covers in the parameter space in Figure B.13. Let us focus on the results for the rAVS model first. Remember that the rAVS model “intuitively” predicts lower ratings for the thin rectangle compared to the tall rectangle (this would be coded as “-”) and no differences in ratings for LOs to the left or to the right of the center-of-mass for the asymmetrical RO (coded as a “0”). This predicted pattern “-0” is produced in almost 80% of the cases for the equality threshold $t_e = 0.5$ (see Figure B.13b) and in almost 95% of the cases for $t_e = 0.1$ (see Figure B.13a). The rAVS model generates only one other pattern, the pattern “00”. This second pattern means that the rAVS model is also able to compute equal ratings for the relative distance condition. This equality can be explained by the role of the rAVS’s α parameter: The smaller the value of α, the smaller the influence of the relative distance. A small α leads to equal ratings, if the influence of the relative distance is smaller than the chosen equality threshold t_e. Accordingly, the volume of this pattern increases with an increasing threshold of equality t_e which is reasonable as ratings that are considered different for a lower threshold are considered equal for a higher threshold.



(a). PSP results for equality of rating $t_e = 0.1$.

(b). PSP results for equality of rating $t_e = 0.5$.

Fig. B.13. Results of the PSP analysis: Relative volume estimates of qualitatively different rating patterns in the parameter space. The first digit codes the difference in ratings for LOs above the thin rectangle compared to LOs above the tall rectangle, the second digit codes for differences in ratings for LOs to the left of the center-of-mass of the L-shaped RO compared to LOs to the right of the center-of-mass (see Figure 2). Two mean ratings were considered equal if they differed less than (a) $t_e = 0.1$ or (b) $t_e = 0.5$. The means of three PSP runs are plotted.

Our previous discussion about the intuitive predictions generated by the AVS model can be summarized in the code “?+”: an unclear prediction for the relative distance test case and higher ratings for LOs that are on the side of the center-of-mass where more mass is located (the left side for the RO in Figure 2b). Interestingly, the PSP algorithm does not reveal any parameter set that enables the AVS model to generate a pattern of ratings that we had expected (no “+” as second digit). In contrast, the AVS model is able to compute the opposite rating pattern for the asymmetrical RO: a higher rating for the LO that is on the side of the center-of-mass where the cavity is located (i.e., to the right for our stimulus). This pattern (a “-” as second digit, i.e., patterns “- -” and “0-”) emerges for more than 65% of the parameter sets for the lower threshold $t_e = 0.1$ (see Figure B.13a). However, it almost disappears for $t_e = 0.5$ (less than 4%; see Figure B.13b). This suggests that the AVS predicts only a slightly higher rating (less than 0.1) for the right LO (above the cavity of the RO) compared to the left LO.

Due to the flexibility of the vector sum in the AVS model, it was unclear what the “intuitive” prediction for the relative distance test case was. The results of the PSP analysis provide evidence for a prediction that overlaps with the prediction from the rAVS model: lower ratings for LOs above the thin rectangle compared to LOs above the tall rectangle (patterns “--” and “-0”). In contrast to the rAVS model, parameter sets that predict no difference for the two test cases (pattern “00”) occupy more volume in the parameter space of the AVS model.

Taken together, the AVS model is more flexible than the rAVS model because it generates a larger number of distinct patterns (four against two generated patterns out of $3^2 = 9$ theoretically possible patterns for $t_e = 0.1$ and $t_e = 0.5$). Moreover, the rAVS model explicitly uses relative distance in its formulation which makes it easier to reason with the model in an intuitive way. The AVS model, on the other hand, makes the same predictions (to some extent) without explicitly using relative distance. One explanation could be that the vector sum implicitly incorporates the use of relative distance.

Appendix B.2.2. PSP Results: AVS-BB and rAVS-CoO Models

We conducted a second PSP analysis with the same settings as before (see Appendix B.2) for our newly proposed models AVS-BB and rAVS-CoO. Since we designed these new models to consider the center-of-object orientation instead of the center-of-mass orientation, we expected them to rate the LOs that are more central with respect to the center-of-object of the L-shaped RO (columns C4–C5, see Figure 8 on page 12) higher than the LOs that are less central but that have equal center-of-mass orientations (columns C2–C3). This corresponds to a negative second digit in the two-digit code that we introduced in the earlier PSP analysis. Looking at the PSP results for the AVS-BB and the rAVS-CoO models shown in Figure B.13, we find this expectation confirmed. For the rating equality threshold $t_e = 0.1$ (Figure B.13a), both models generate such rating patterns (“--” and “0-”) for more than 80% of their parameter spaces. For $t_e = 0.5$ (Figure B.13b), the volumes of these patterns decrease but still cover more than 60% of their parameter spaces.

Next, we note that the AVS-BB model is slightly more flexible than the rAVS-CoO model in terms of the number of generated data patterns: Out of $3^2 = 9$ possible data patterns, the AVS-BB model generates four while the rAVS-CoO model only generates three patterns. However, the additional data pattern generated by the AVS-BB model (“0-”) is the empirical rating pattern (no difference in ratings across rectangular ROs and higher ratings for more central LOs compared to less central LOs above asymmetrical ROs). While this favors the AVS-BB model over the rAVS-CoO model (which does not generate the empirical pattern at all), we note that (i) the AVS-BB model generates the empirical pattern not as a main prediction (it is generated with comparably few parameter sets only: ca. 6% of the parameter space for $t_e = 0.1$ and ca. 15% for $t_e = 0.5$) and (ii) the mechanisms of the AVS-BB model cannot explain all details of our empirical findings regarding the influence of the relative distance. On the other hand, the rAVS-CoO model clearly does not capture the influence of relative distance appropriately due to the prior misconception of this effect in the rAVS model (and we did not aim to refine any model to account for the empirical findings). This explains why the rAVS-CoO model does not generate the empirical pattern “0-”.

Appendix B.3. Model Flexibility Analysis

The MFA computes an intuitively graspable ratio ϕ of the number of model predictions to the number of thinkable data patterns: $\phi = \frac{\text{number of different model outputs}}{\text{number of all possible data patterns}}$. A model with a high MFA ratio can be used to predict almost every observable data. Such a model is hard to falsify and useless in terms of explaining cognition (e.g., it might predict more implausible than plausible data). A model with a low ratio, on the other hand, tightly constrains the space of predicted data. Such a model makes strong claims about the task as it cannot produce a wide range of data that is theoretically possible.

Method. The MFA computes the ratio ϕ by enumerating the whole parameter space of a model, i.e., all possible parameter settings. Having computed all possible model outputs, the MFA determines the area of the space covered by these predictions with respect to all possible data patterns. To do so, every dimension of the data space is split into intervals. Veksler et al. (2015) suggest to use $\sqrt[n]{j^4}$ intervals where n is the dimension of one data pattern (448 for our whole stimuli set) and j the number of intervals per parameter (we chose $j = 50$). If two model predictions fall into the same interval across all dimensions of the data patterns, they are considered equal. The number of unequal model predictions is counted and divided by the number of all intervals (which equals the number of all computed predictions $\left(\sqrt[n]{50^4}\right)^n = 50^4$).

The suggestion by Veksler et al. (2015) to split every dimension of the data space into $\sqrt[n]{50^4}$ leads to approximately 1.04 intervals (for our whole stimuli set). Remember that every dimension of the data space ranges from 1 to 9 (the rating scale). Splitting one dimension into 1.04 intervals means that all ratings from 1 to 8.64 are considered equal (they fall into the first interval) and all ratings greater than 8.64 are considered different (they fall into the second interval). Arguably, this number of intervals is not reasonable because it implies that most ratings are considered to be equal. To address this problem, we did not follow the $\sqrt[n]{j^4}$ suggestion by Veksler et al. (2015) but set the number of intervals to a domain-specific value (in our case: the range of the rating scale). However, note that by doing so we define more intervals than we have model predictions. This means that a ϕ ratio of 1.0 can never occur. In order to account for this, we report below the normalized $\phi_n = \frac{\phi}{\phi_{max}}$ by dividing with the highest possible ϕ

value (that depends on the dimensions of the stimuli set): $\phi_{max} = 50^4/ratingRange^{dataSpaceDimension}$.

For the parameters, we used the same range as for the other simulation methods (see equations B.3–B.7). Considering a recent critique that reported invariances of the MFA for different parameter ranges (Evans, Howard, Heathcote, & Brown, 2017), we computed all MFA results a second time using smaller but still plausible parameter ranges for some parameters. With one exception, we found higher flexibilities for these smaller parameter ranges compared to the MFA results for the greater parameter range. Accordingly, the *absolute* ϕ value should not be interpreted because seemingly unrelated changes in its computation end up in large absolute differences. However, the relative rankings of the models did only change in one case. This is why we think it is still legitimate to discuss the relative differences in flexibilities as estimated by the MFA – despite the critique from Evans et al. (2017). Nevertheless, the MFA results should be interpreted with caution and related with the outcomes of other methods that provide different perspectives on model flexibility (e.g., PSP or landscaping).

The flexibility of a model varies on the stimuli used as input. This is why we computed all MFA ϕs for each model on four different subsets: our whole stimuli set, only the rectangular ROs, only the asymmetrical ROs, and the whole stimuli set from Regier and Carlson (2001).

Results. Figure B.14 plots the MFA ϕ_n values. The AVS model is the most flexible model across almost all stimuli sets (except for the subset of our stimuli that consists of the rectangular ROs). The second most flexible model is the rAVS model. This is consistent with the PSP results reported earlier, where the AVS model generated more patterns than the rAVS model. A potential reason for the higher flexibility of the AVS model compared to the rAVS model is the use of a *population* of vectors. Conceptually, the rAVS model also uses a population of vectors, however, de facto, the rAVS model needs to compute only a single vector for all considered stimuli. This is because the LO on which the vector sum is rooted in the rAVS model consists of only one single point (see Kluth, Burigo, & Knoeferle, 2017, for a more elaborated discussion on this difference between the AVS model and the rAVS model). However, since (i) our recomputation of the MFA results using a smaller parameter range changed the relative ordering of the AVS and the rAVS model for our whole stimuli set[14], (ii) the rAVS model is also more flexible for the rectangular ROs and (iii) one should exercise caution in interpreting MFA results (Evans et al., 2017), our MFA results cannot reliably distinguish the AVS and the rAVS model in terms of their flexibility.

Of greater interest in light of our surprising empirical finding that the center-of-object seems to be of greater importance for spatial language evaluation compared to the center-of-mass is the fact that both models implementing this finding (AVS-BB and rAVS-CoO) are less flexible than their corresponding origins (AVS and rAVS). While this advantage is less pronounced for the stimuli from Regier and Carlson (2001), it is stronger for our whole stimuli set and strongest for the subset with the asymmetrical ROs – the data set which motivated the development of the AVS-BB model and the rAVS-CoO model in the first place.[15]

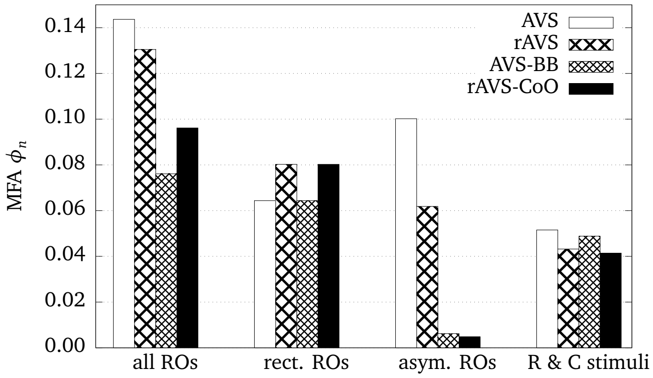


Fig. B.14. Normalized results of the Model Flexibility Analysis (MFA) computed with 50^4 model predictions and as many intervals for every dimension of the data space as there were rating intervals (i.e., 9 for our stimuli, 10 for stimuli from Regier & Carlson, 2001).

How do the rAVS-CoO model and the AVS-BB model compare to each other with respect to their flexibility? The AVS-BB model is less flexible for our whole stimuli set but slightly more flexible for the asymmetrical ROs and the stimuli from Regier and Carlson (2001). Due to these conflicting differences (that are also small in the magnitude) and the caution one should use in interpreting MFA results (Evans et al., 2017), our MFA simulation results do favor neither the AVS-BB nor the rAVS-CoO model.

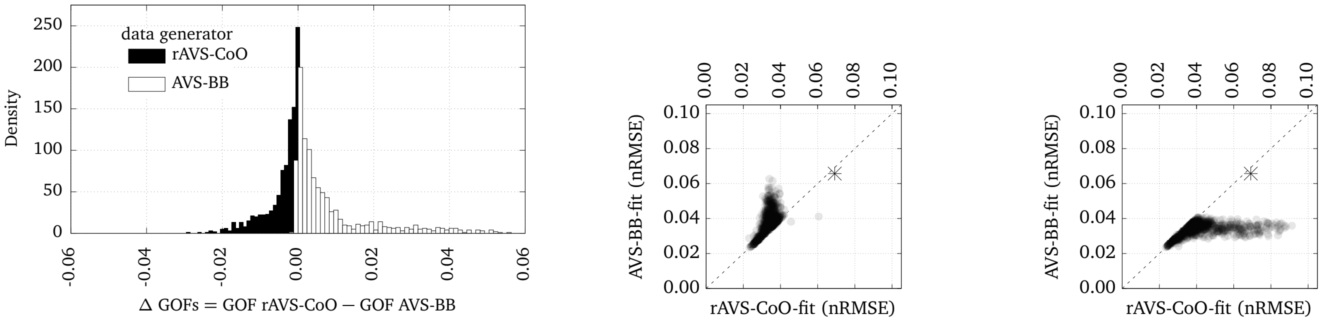
The modified models (AVS-BB and rAVS-CoO) are less flexible than the models they originated from (AVS and rAVS). Together with the better GOF and SHO results for the AVS-BB and the rAVS-CoO model (compared to the AVS and rAVS model), this lower flexibility further supports the overall superior performance of the newly proposed models. However, we cannot reliably distinguish the models that incorporate a directionality of attention from the RO to the LO (i.e., AVS and AVS-BB) from the models that incorporate the reversed directionality from the LO to the RO (i.e., rAVS and rAVS-CoO) in terms of their model flexibility.

Appendix B.4. Landscaping

Method. The main idea of landscaping is the following: Given model input (i.e., ROs and LOs in our case), each model is used to generate sets of artificial data (i.e., ratings in our case)

[14] In general, the rAVS model showed a greater increase in flexibility than the AVS model in these recomputations. A very likely reason for this is that the α parameter in the rAVS model has a stronger influence on model flexibility than the λ parameter in the AVS model.

[15] For the rectangular ROs, the two new models have the exact same flexibility as the AVS and the rAVS model. This is because for the rectangular ROs the center-of-object coincides with the center-of-mass of the RO and hence the AVS-BB model equals the AVS model and the rAVS-CoO model equals the rAVS model.



(a). Histograms of differences in fits to artificial data.

(b). Model fits to artificial data generated by the rAVS-CoO model.

(c). Model fits to artificial data generated by the AVS-BB model.

Fig. B.15. Landscaping results contrasting the rAVS-CoO model with the AVS-BB model on the asymmetrical ROs (collapsing across *über*, *above*, and *unter*, *below*). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 9c).

and then both models fit these data. The data-uninformed version of the parametric bootstrap cross-fitting method (PBCM, Wagenmakers et al., 2004) is the same as the landscaping method except for the fact that in the PBCM no noise is added to the generated data and the results are plotted in a different way. We plot the results of the landscaping analysis in both ways (histograms for PBCM, landscaping plots for landscaping) but follow the landscaping procedure described in Navarro et al. (2004) which we will briefly explain in the following paragraph.

To generate artificial data, model parameters are randomly chosen and the output of each model with these parameters for the given input is computed. We sampled each parameter from a uniform distribution over the corresponding entire parameter range (see equations B.3–B.7). Next, noise is added to the generated data before both models fit the data. We added Gaussian noise with a standard deviation of 0.3 to each generated rating. The magnitude of this noise stems from the size of the standard error of the mean for our whole data set. This procedure is repeated several times with both models acting as the data generating model. We generated 1000 sets of artificial data from each model. Applying the landscaping method, we contrasted the following model pairs: (i) the two best models (AVS-BB and rAVS-CoO) using our asymmetrical ROs and the stimuli from Regier and Carlson (2001) as input, (ii) the rAVS model and the rAVS-CoO model using our asymmetrical ROs and our whole stimuli set and (iii) the rAVS model and the AVS model on the stimuli from Regier and Carlson (2001).

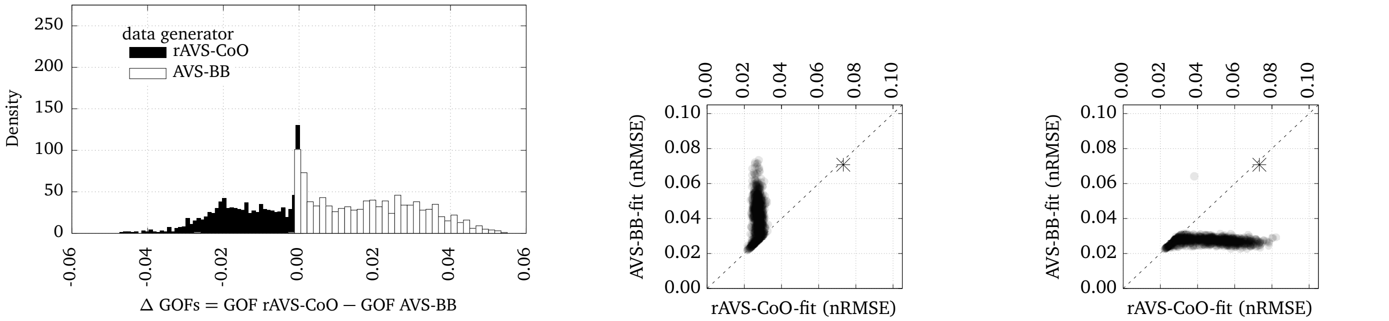
Appendix B.4.1. Landscaping Results

AVS-BB Versus rAVS-CoO: Asymmetrical ROs. The landscaping results are shown in Figure B.15. Looking at the landscape plots (Figure B.15b and B.15c) reveals that the data generating model mostly fitted the data better than the other model as is evident by the location of the model fits on only one side of the dashed diagonal line of equal fit. The landscape plots also contain the fits to the empirical data (asterisks) which were already shown in Figure 9c (the GOF bars for AVS-BB and rAVS-CoO). These fits are of equal magnitude compared to the fits of the model that did not generate the data suggesting that in general the models produce data close to the empirical data.[16] The fits to empirical data are, however, slightly worse than the fits of the data generating model to the artificial data. This could indicate that either (i) the empirical data contain more noise than we added to the artificial data or (ii) that the artificial data are reliably different from the empirical data (although this difference is small, see magnitude of fits). Both reasons would explain why we could not yet distinguish the models on the empirical data.

Although the fits to artificial data in Figure B.15b and Figure B.15c are plotted with transparency so that areas with higher density appear darker, the density distribution is hardly visible due to the large number of fits. Therefore, Figure B.15a combines both landscape plots and depicts histograms of differences in GOF (the proposed plot for the PBCM, Wagenmakers et al., 2004). Negative values code better GOFs (lower nRMSEs) for rAVS-CoO compared to AVS-BB; positive values code better GOFs for AVS-BB compared to rAVS-CoO. The filled bars originate from fits to data generated by the rAVS-CoO model (corresponding to Figure B.15b), the empty bars depict Δ GOFs for data generated by the AVS-BB model (corresponding to Figure B.15c).

The histograms in Figure B.15a peak around 0.0. This means that a considerable number of model fits are practically equivalent (i.e., they lie very close by the line of equal fit in the landscape plots). Equal fits can be interpreted as a symptom of model mimicry: Although model A generated the data, model B is able to closely reproduce it by mimicking model A. Two models mimicking each other would result in fits along the line of equal fit in the landscape plot and a histogram peaking sharply around 0.0. However, the tails of each histogram only go in one direction, reflecting a considerable amount of data that is better fitted by the data-generating model. This is supported by the shape of the landscape plots: Although the fits are somewhat attached to the line of equal

[16] Note, however, that the comparison of empirical to "typical" model data is only done by comparing the *fits* of the models to these data sets. In principle, one could think of two completely different data sets that can be fitted equally well by the same model.



(a). Histograms of differences in fits to artificial data.

(b). Model fits to artificial data generated by the rAVS-CoO model.

(c). Model fits to artificial data generated by the AVS-BB model.

Fig. B.16. Landscaping results contrasting the rAVS-CoO model with the AVS-BB model on the stimuli from Regier and Carlson (2001). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 9d).

fit in Figures B.15b and B.15c, they do not completely follow it. This is to say that the two models do not fully mimic each other on our asymmetrical ROs but that they still are able to closely fit the not self-generated data (see overall low magnitude of model fits).

However, the AVS-BB model shows a higher degree of model mimicry than the rAVS-CoO, as can be seen by comparing the shape of the landscape for the data generated by the rAVS-CoO model (Figure B.15b) with the landscape shape for the data generated by the AVS-BB model (Figure B.15c). The former is more attached to the line of equal fit than the latter. Figure B.15a shows the same trend with a smaller filled than empty tail. That is, the AVS-BB model fits the data generated by the rAVS-CoO model almost as well as the rAVS-CoO model itself while the rAVS-CoO model shows a worse (but still good) performance for the data generated by the AVS-BB model. This trend exists to a lesser extent also for the stimuli from Regier and Carlson (2001, see Figure B.16).

AVS-BB Versus rAVS-CoO: Stimuli from Regier and Carlson (2001). Compared to the landcaping analysis of the AVS-BB and the rAVS-CoO model for the asymmetrical ROs, the degree of model mimicry is lower for the stimuli from Regier and Carlson (2001). This is reflected in only a decent peak around 0.0 in the histogram (Figure B.16a) and the shapes of the landscapes (Figures B.16b and B.16c) that are more orthogonal to the axis plotting the fit of the data generating model (compared to Figure B.15). Even though, there is a small hint that the AVS-BB model mimics the rAVS-CoO model more than vice versa: The AVS-BB model obtains slightly better fits to the data generated by the rAVS-CoO model than the rAVS-CoO model does on the data generated by the AVS-BB model. This is reflected in a longer tail of the histogram with the empty bars compared to the histogram with the filled bars (Figure B.16a) and in the longer landscape in Figure B.16c than in Figure B.16b.

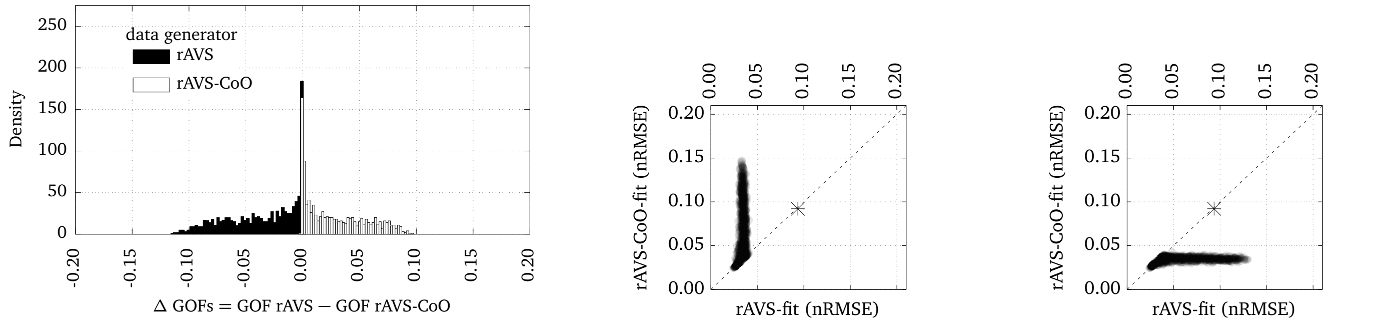
rAVS Versus rAVS-CoO: Whole Stimuli Set and Asymmetrical ROs. We applied the landscaping analysis for the rAVS and the rAVS-CoO model to investigate the indistinguishable model fitting performance of the rAVS and the rAVS-CoO model on our whole data set (see Figure 9a); by contrast, for the rating data for the asymmetrical ROs, the rAVS-CoO model outperformed the rAVS model (see Figure 9c). One reason for this contrasting outcome might be that the parameter sets that work well on the asymmetrical ROs for the rAVS-CoO model are not a good choice for the whole data set.

The results of the landscaping analysis are shown in Figures B.17 and B.18. Looking at the landscape plots (panels (b) and (c)) reveals that the data generating model mostly fitted the data better than the other model as is evident by the location of the model fits on only one side of the dashed diagonal line of equal fit. Since the spread of the model fits is orthogonally aligned with the axis denoting the fit of the data generating model and the histograms have long, flat tails, the two models do not mimic each other.

However, the rAVS model produces overall slightly better fits to not-self-generated data compared to the rAVS-CoO model as is evident in the landscape plots by the smaller horizontal spread of model fits in panels (c) compared to the vertical spread of model fits in panels (b). The same effect is visible in the histograms by smaller tails for the empty bars compared to the filled bars. This effect is more pronounced if only the asymmetrical ROs are used as stimuli (Figure B.18) compared to using all ROs (Figure B.17). Apparently, fitting certain rating patterns from the rAVS model is more problematic for the rAVS-CoO model than fitting rating patterns generated by the rAVS-CoO model is for the rAVS model. This makes the rAVS model slightly more flexible than the rAVS-CoO model – in line with the results from the MFA.

To sum up the comparison of the rAVS and the rAVS-CoO model on the two data sets, one can say that the equal performance on the whole data set is not due to model mimicry. In contrast, the rAVS-CoO model could not fit the whole data set as well as the asymmetrical ROs data set possibly because it is not flexible enough for these parts of the data space. A potential reason for this missing flexibility is the failure to properly account for the effects of relative distance.

rAVS Versus AVS: Stimuli from Regier and Carlson (2001). We analyzed the similar performance of the rAVS and the AVS

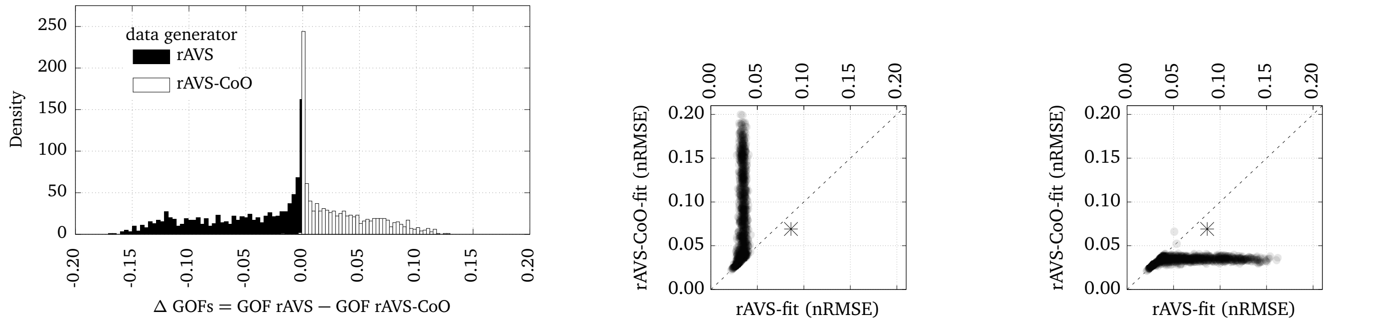


(a). Histograms of differences in fits to artificial data.

(b). Model fits to artificial data generated by the rAVS model.

(c). Model fits to artificial data generated by the rAVS-CoO model.

Fig. B.17. Landscaping results contrasting the rAVS model with the rAVS-CoO model on our whole stimuli set (collapsing across *über*, *above*, and *unter*, *below*). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 9a).



(a). Histograms of differences in fits to artificial data.

(b). Model fits to artificial data generated by the rAVS model.

(c). Model fits to artificial data generated by the rAVS-CoO model.

Fig. B.18. Landscaping results contrasting the rAVS model with the rAVS-CoO model on the asymmetrical ROs only (collapsing across *über*, *above*, and *unter*, *below*). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 9c).

model on the data from Regier and Carlson (2001) using the landscaping method. On these data, the AVS model is more flexible than the rAVS model, as shown with the Model Flexibility Analysis (see Kluth, Burigo, & Knoeferle, 2017, or Section Appendix B.3). Here, we examine whether this difference in model flexibility is also reflected in the landscaping method. The landscaping results are plotted in Figure B.19.

These results do look very similar to the results from the landscaping analysis that contrasted the rAVS-CoO model with the AVS-BB model on the stimuli from Regier & Carlson, 2001 (discussed above and shown in Figure B.16). The fits with the unmodified models (AVS and rAVS), however, are slightly better than those for the modified models (rAVS-CoO and AVS-BB). Also, the AVS and the rAVS model have an even lower degree of model mimicry (comparing the peaks of histograms).

The higher flexibility of the AVS model is visible in Figure B.19 in (i) the histogram plot (larger tail for the empty bars compared to the filled bars) and (ii) in the landscape plots (larger horizontal spread for data generated by the AVS model, panel (c), than vertical spread for data generated by the rAVS model, panel (b)). Note that the degree of this difference in model flexibility is very small corresponding to the small MFA values on this data set reported in Kluth, Burigo, and Knoeferle (2017) and shown in Figure B.14.

References

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica*, *137*(2), 181–189. doi: 10.1016/j.actpsy.2010.09.008

Burigo, M., & Coventry, K. (2005). Reference frame conflict in assigning direction to space. In C. Freksa, M. Knauff, B. Krieg Brückner, B. Nebel, & T. Barkowsky (Eds.), *Spatial Cognition IV. Reasoning, Action, Interaction* (Vol. 3343, pp. 111–123). Berlin: Springer.

Burigo, M., Coventry, K. R., Cangelosi, A., & Lynott, D. (2016). Spatial language and converseness. *Quarterly Journal of Experimental Psychology*, *69*(12), 2319-2337. doi: 10.1080/17470218.2015.1124894

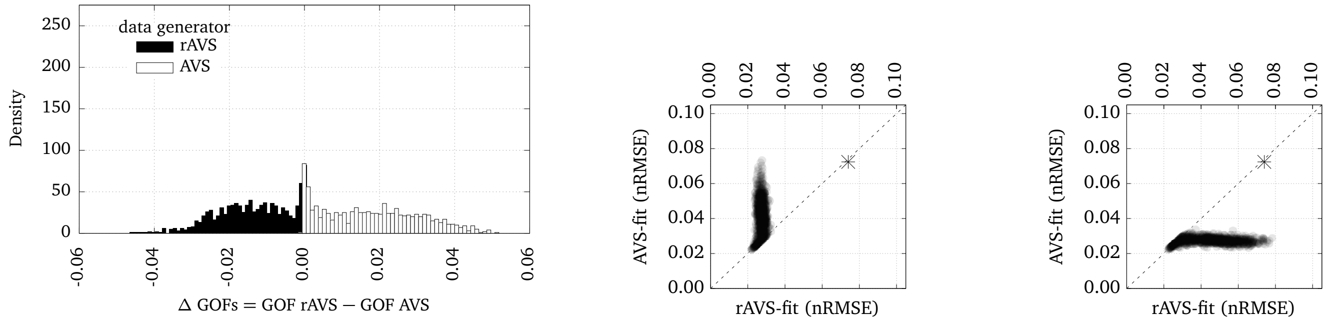
Burigo, M., & Knoeferle, P. (2015). Visual attention during spatial language comprehension. *PLoS ONE*, *10*(1), e0115758. doi: 10.1371/journal.pone.0115758

Burigo, M., & Sacchi, S. (2013). Object orientation affects spatial language comprehension. *Cognitive Science*, *37*(8), 1471–1492. doi: 10.1111/cogs.12041

Burigo, M., & Schultheis, H. (2018). The effects of direction and orientation of located objects on spatial language comprehension. *Language & Cognition*, *10*(2), 298–328. doi: 10.1017/langcog.2018.3

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Canty, A., & Ripley, B. (2016). boot: Bootstrap R (S-Plus) Functions [Computer software manual]. (R package version 1.3-18)



(a). Histograms of differences in fits to artificial data.

(b). Model fits to artificial data generated by the rAVS model.

(c). Model fits to artificial data generated by the AVS model.

Fig. B.19. Landscaping results contrasting the rAVS model with the AVS model on the stimuli from Regier and Carlson (2001). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 9d).

Carlson, L. A. (2003). Using spatial language. *Psychology of Learning and Motivation*, *43*, 127–162.

Carlson, L. A., & Logan, G. D. (2001). Using spatial terms to select an object. *Memory & 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). Elsevier.

Carlson, L. A., Regier, T., Lopez, W., & Corrigan, B. (2006). Attention unites form and function in spatial language. *Spatial Cognition and Computation*, *6*(4), 295–308.

Carlson, L. A., & Van Deman, S. R. (2004). The space in spatial language. *Journal of Memory and Language*, *51*(3), 418–436. doi: 10.1016/j.jml.2004.06.004

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.

Carlson-Radvansky, L. A., & Logan, G. D. (1997). The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, *37*(3), 411–437.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*(1), 30–49.

Cohen, E. H., Schnitzer, B. S., Gersch, T. M., Singh, M., & Kowler, E. (2007). The relationship between spatial pooling and attention in saccadic and perceptual tasks. *Vision Research*, *47*(14), 1907–1923. doi: 10.1016/j.visres.2007.03.018

Conder, J., Fridriksson, J., Baylis, G. C., Smith, C. M., Boiteau, T. W., & Almor, A. (2017). Bilateral parietal contributions to spatial language. *Brain and Language*, *164*, 16–24. doi: 10.1016/j.bandl.2016.09.007

Coventry, K. R., Lynott, D., Cangelosi, A., Monrouxe, L., Joyce, D., & Richardson, D. C. (2010). Spatial language, visual attention, and perceptual simulation. *Brain and Language*, *112*(3), 202–213. doi: 10.1016/j.bandl.2009.06.001

Coventry, K. R., Prat Sala, M., & Richards, L. (2001). The interplay between geometry and function in the comprehension of *over*, *under*, *above*, and *below*. *Journal of Memory and Language*, *44*(3), 376–398. doi: 10.1006/jmla.2000.2742

Crawford, L. E., Regier, T., & Huttenlocher, J. (2000). Linguistic and non-linguistic spatial categorization. *Cognition*, *75*(3), 209–235.

[dataset]Kluth, T. (2018). *A* C++ *implementation of cognitive models of spatial language understanding as well as pertinent empirical data and analyses.* Bielefeld University. doi: 10.4119/unibi/2918231

Dawson, M. R. (1988). Fitting the ex-Gaussian equation to reaction time distributions. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 54–57.

Desanghere, L., & Marotta, J. J. (2015). The influence of object shape and center of mass on grasp and gaze. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01537

Dessalegn, B., & Landau, B. (2008). More than meets the eye: The role of language in binding and maintaining feature conjunctions. *Psychological Science*, *19*(2), 189–195.

Dessalegn, B., & Landau, B. (2013). Interaction between language and vision: It’s momentary, abstract, and it develops. *Cognition*, *127*, 331–344. doi: doi:10.1016/j.cognition.2013.02.003

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. doi: 10.1177/1745691611406920

Eaton, J. W., Bateman, D., Huauberg, S., & Wehbring, R. (2015). *GNU Octave version 4.0.0 manual: A high-level interactive language for numerical computations*. (http://www.gnu.org/software/octave/doc/interpreter)

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219–234. doi: 10.3758/s13423-017-1317-5

Evans, N. J., Howard, Z. L., Heathcote, A., & Brown, S. D. (2017). Model flexibility analysis does not measure the persuasiveness of a fit. *Psychological Review*, *124*(3), 339. doi: 10.1037/rev0000057

Fernandez-Duque, D., & Johnson, M. L. (1999). Attention metaphors: How metaphors guide the cognitive psychology of attention. *Cognitive Science*, *23*(1), 83–116.

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. doi: 10.1016/j.cognition.2011.11.002

Gabry, J. (2017). *bayesplot: Plotting for Bayesian Models.* http://mc-stan.org/. (R package version 1.2.0)

Galassi, M., Davies, J., Theiler, J., Gough, J., Jungman, G., Alken, P., … Rossi, F. (2009). *GNU Scientific Library Reference Manual* (3rd ed.). Network Theory Ltd. (https://www.gnu.org/software/gsl)

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. doi: 10.1007/s11222-013-9416-2

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *233*(4771), 1416-1419.

Gibson, B. S., & Kingstone, A. (2006). Visual attention and the semantics of space beyond central and peripheral cues. *Psychological Science*, *17*(7), 622–627.

Gibson, B. S., & Sztybel, P. (2014). The spatial semantics of symbolic attention control. *Current Directions in Psychological Science*, *23*(4), 271–276. doi: 10.1177/0963721414536728

Gibson, B. S., Thompson, A. N., Davis, G. J., & Biggs, A. T. (2011). Going the distance: Extra-symbolic contributions to the symbolic control of spatial attention. *Visual Cognition*, *19*(10), 1237–1261. doi: 10.1080/13506285.2011.628636

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. doi: 10.1016/j.socec.2004.09.033

Hayward, W. G., & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, *55*(1), 39–84.

Hayworth, K. J., Lescroart, M. D., & Biederman, I. (2011). Neural encoding of relative position. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1032. doi: 10.1037/a0022338

Holcombe, A. O., Linares, D., & Vaziri-Pashkam, M. (2011). Perceiving spatial relations via attentional tracking and shifting. *Current Biology*, *21*(13), 1135–1139. doi: 10.1016/j.cub.2011.05.031

Hörberg, T. (2008). Influences of form and function on the acceptability of projective prepositions in Swedish. *Spatial Cognition & Computation*, *8*(3), 193–218. doi: 10.1080/13875860801993652

Huttenlocher, J., & Strauss, S. (1968). Comprehension and a statement's relation to the situation it describes. *Journal of Verbal Learning and Verbal Behavior*, *7*(2), 300–304.

Kim, W., Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). An MCMC-based method of comparing connectionist models in cognitive science. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems 16: Proceedings of the 2003 conference* (pp. 937–944). Canada: MIT Press.

Kluth, T. (in preparation). *Modeling the contribution of visual attention to spatial language verification* (Unpublished doctoral dissertation). Cognitive Interaction Technology Excellence Cluster (CITEC) Graduate School, Bielefeld University. (to be published under https://pub.uni-bielefeld.de/person/54885831/)

Kluth, T., Burigo, M., & Knoeferle, P. (2016a). Investigating the parameter space of cognitive models of spatial language comprehension. In *5. interdisziplinärer Workshop Kognitive Systeme: Mensch, Teams, Systeme und Automaten. Verstehen, Beschreiben und Gestalten Kognitiver (Technischer) Systeme.* Bochum, Germany.

Kluth, T., Burigo, M., & Knoeferle, P. (2016b). Modeling shifts of attention during spatial language comprehension. In T. Tenbrink, A. Foltz, A. Wallington, J. O. Redondo, J. Ryan, & E. Bedford (Eds.), *UK-CLC 2016 Conference Proceedings* (p. 71). Bangor, Wales, UK.

Kluth, T., Burigo, M., & Knoeferle, P. (2017). Modeling the directionality of attention during spatial language comprehension. In J. v. d. Herik & J. Filipe (Eds.), *Agents and Artificial Intelligence* (Vol. 10162, pp. 283–301). Springer International Publishing AG. doi: 10.1007/978-3-319-53354-4_16

Kluth, T., Burigo, M., Schultheis, H., & Knoeferle, P. (2016a). Distinguishing cognitive models of spatial language understanding. In D. Reitter & F. E. Ritter (Eds.), *Proceedings of the International Conference on Cognitive Modeling* (pp. 230–231). University Park, Pennsylvania, USA: Penn State.

Kluth, T., Burigo, M., Schultheis, H., & Knoeferle, P. (2016b). The role of the center-of-mass in evaluating spatial language. In T. Barkowsky, H. Schultheis, J. van de Ven, & Z. Falomir Llansola (Eds.), *13th Biannual Conference of the German Society for Cognitive Science: Proceedings* (pp. 11–14). Bremen, Germany.

Kluth, T., Burigo, M., Schultheis, H., & Knoeferle, P. (2017). Size matters: Effects of relative distance on the acceptability of spatial prepositions. In A. Shestakova et al. (Eds.), *Proceedings of the 10th Embodied and Situated Language Processing Conference* (p. 21). Moscow, Russia: Centre for Cognition and Decision Making, Higher School of Economics.

Kluth, T., & Schultheis, H. (2014). Attentional distribution and spatial language. In C. Freksa, B. Nebel, M. Hegarty, & T. Barkowsky (Eds.), *Spatial Cognition IX* (Vol. 8684, pp. 76–91). Springer International Publishing. doi: 10.1007/978-3-319-11215-2_6

Kluth, T., & Schultheis, H. (2018). Rating distributions and Bayesian inference: Enhancing cognitive models of spatial language use. In M. Idiart, A. Lenci, T. Poibeau, & A. Villavicencio (Eds.), *Proceedings of the Eighth Workshop on Cognitive Aspects of Computational Language Learning and Processing, co-located with the 56th Annual Meeting of the Association for Computational Linguistics.* Melbourne, Australia: Association for Computational Linguistics.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573. doi: 10.1037/a0029146

Landau, B. (2017). Update on "What" and "Where" in Spatial Language: A New Division of Labor for Spatial Terms. *Cognitive Science*, *41*(S2), 321–350. doi: 10.1111/cogs.12410

Landau, B., & Jackendoff, R. (1993). Whence and whither in spatial language and spatial cognition? *Behavioral and Brain Sciences*, *16*(02), 255–265.

Lee, C., Rohrer, W. H., & Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, *332*, 357-360.

Levinson, S. C. (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. (preprint, retrieved from osf.io/9h3et) doi: 10.17605/OSF.IO/9H3ET

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*(1), 22–25.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490. doi: 10.1037/a0022643

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

Logan, G. D. (1995). Linguistic and conceptual control of visual spatial attention. *Cognitive Psychology*, *28*(2), 103–174.

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

Logan, G. D., & Zbrodoff, N. J. (1999). Selection for cognition: Cognitive constraints on visual spatial attention. *Visual Cognition*, *6*(1), 55–81.

Lovett, A., & Forbus, K. (2009). Using a Visual Routine to Model the Computation of Positional Relationships. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1882–1887). Austin, TX, USA: Cognitive Science Society.

Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3), 108–114.

Matin, E., Shao, K., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Attention, Perception, & Psychophysics*, *53*(4), 372–380.

Melcher, D., & Kowler, E. (1999). Shapes, surfaces and saccades. *Vision Research*, *39*(17), 2929–2946.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Molenberghs, P., Mesulam, M. M., Peeters, R., & Vandenberghe, R. R. (2007). Remapping attentional priorities: differential contribution of superior parietal lobule and intraparietal sulcus. *Cerebral Cortex*, *17*(11), 2703–2712. doi: 10.1093/cercor/bhl179

Navarro, D. J., Myung, I. J., Pitt, M. A., & Kim, W. (2003). Global model analysis by landscaping. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society.* Cognitive Science Society.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*(1), 47–84. doi: 10.1016/j.cogpsych.2003.11.001

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8), 1–19. doi: 10.1167/10.8.20

O'Keefe, J. (2003). Vector grammar, places, and the functional role of the spatial prepositions in English. In E. van der Zee & J. Slack (Eds.), *Representing Direction in Language and Space* (pp. 69–85). Oxford University Press.

Pederson, E. (2003). How many reference frames? In C. Freksa, W. Brauer, C. Habel, & K. F. Wender (Eds.), *Spatial Cognition III – Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Reasoning* (Vol. 2685, pp. 287–304). Berlin: Springer.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57–83. doi: 10.1037/0033-295X.113.1.57

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in*

Cognitive Sciences, *6*(10), 421–425.

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*(1), 127–158.

R Core Team. (2016). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. (https://www.R-project.org/)

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

Regier, T., & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, *130*(2), 273–298. doi: 10.1037//0096-3445.130.2.273

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, USA: Cognitive Science Society.

Richter, M., Lins, J., & Schöner, G. (2016). A Neural Dynamic Model Parses Object-Oriented Actions. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society.* Austin, TX, USA: Cognitive Science Society.

Richter, M., Lins, J., & Schöner, G. (2017). A Neural Dynamic Model Generates Descriptions of Object-Oriented Actions. *Topics in Cognitive Science*, *9*(1), 35–47. doi: 10.1111/tops.12240

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367. doi: 10.1037//0033-295X.107.2.358

Roth, J. C., & Franconeri, S. L. (2012). Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in Psychology*, *3*(464). doi: 10.3389/fpsyg.2012.00464

Roy, D., & Mukherjee, N. (2005). Towards situated speech understanding: Visual context priming of language models. *Computer Speech & Language*, *19*(2), 227–248. doi: 10.1016/j.csl.2004.08.003

Schultheis, H., & Carlson, L. A. (2017). Mechanisms of Reference Frame Selection in Spatial Term Use: Computational and Empirical Studies. *Cognitive Science*, *41*(2), 276–325. doi: 10.1111/cogs.12327

Schultheis, H., & Carlson, L. A. (2018). Inter-process relations in spatial language: Feedback and graded compatibility. *Cognition*, *176*, 140–158.

Schultheis, H., Singhaniya, A., & Chaplot, D. S. (2013). Comparing model comparison methods. In M. Knauff, M. Pauen, & I. Sebanz N. & Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1294–1299). Austin, TX, USA: Cognitive Science Society.

Stan Development Team. (2016). *RStan: The R interface to Stan.* (R package version 2.14.1)

Talmy, L. (2000). *Towards a Cognitive Semantics* (Vol. I: Concept Structuring Systems). MIT Press.

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.

The CGAL Project. (2015). *CGAL User and Reference Manual* (4.7 ed.). CGAL Editorial Board. (https://doc.cgal.org/4.7/Manual/packages.html)

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*(3), 424–465.

Vehtari, A., Gelman, A., & Gabry, J. (2016). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.* https://github.com/stan-dev/loo. (R package version 1.0.0.)

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2015). Model flexibility analysis. *Psychological Review*, *122*(4), 755–769. doi: 10.1037/a0039657

Vishwanath, D., & Kowler, E. (2003). Localization of shapes: Eye movements and perception compared. *Vision Research*, *43*(15), 1637–1653. doi: 10.1016/S0042-6989(03)00168-8

Vishwanath, D., & Kowler, E. (2004). Saccadic localization in the presence of cues to three-dimensional shape. *Journal of Vision*, *4*(6), 4–4.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. doi: 10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*(1), 28–50. doi: 10.1016/j.jmp.2003.11.004

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. (http://ggplot2.org)

Williams, T., Kelley, C., & many others. (2016). *Gnuplot 5.0.5: An interactive plotting program.* http://www.gnuplot.info/.

Wilson, H. R., & Kim, J. (1994). Perceived Motion in the Vector Sum Direction. *Vision Research*, *34*(14), 1835–1842.

Yuan, L., Uttal, D., & Franconeri, S. (2016). Are Categorical Spatial Relations Encoded by Shifting Visual Attention between Objects? *PLoS ONE*, *11*(10), e0163141. doi: 10.1371/journal.pone.0163141