Modeling the Directionality of Attention During Spatial Language Comprehension

Thomas Kluth[1(✉)], Michele Burigo[1], and Pia Knoeferle[2]

[1] Language & Cognition Group, CITEC (Cognitive Interaction Technology Excellence Cluster), Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany
{tkluth,mburigo}@cit-ec.uni-bielefeld.de

[2] Department of German Language and Linguistics, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany
pia.knoeferle@hu-berlin.de

Abstract. It is known that the comprehension of spatial prepositions involves the deployment of visual attention. For example, consider the sentence "The salt is to the left of the stove". Researchers [29,30] have theorized that people must shift their attention from the stove (the reference object, RO) to the salt (the located object, LO) in order to comprehend the sentence. Such a shift was also implicitly assumed in the Attentional Vector Sum (AVS) model by [35], a cognitive model that computes an acceptability rating for a spatial preposition given a display that contains an RO and an LO. However, recent empirical findings showed that a shift from the RO to the LO is not necessary to understand a spatial preposition ([3], see also [15,38]). In contrast, these findings suggest that people perform a shift in the reverse direction (i.e., from the LO to the RO). Thus, we propose the reversed AVS (rAVS) model, a modified version of the AVS model in which attention shifts from the LO to the RO. We assessed the AVS and the rAVS model on the data from [35] using three model simulation methods. Our simulations show that the rAVS model performs as well as the AVS model on these data while it also integrates the recent empirical findings. Moreover, the rAVS model achieves its good performance while being less flexible than the AVS model. (This article is an updated and extended version of the paper [23] presented at the 8th International Conference on Agents and Artificial Intelligence in Rome, Italy. The authors would like to thank Holger Schultheis for helpful discussions about the additional model simulation.)

Keywords: Spatial language · Spatial prepositions · Cognitive modeling · Model flexibility · Visual attention

1 Introduction

Imagine a household robot that helps you in the kitchen. You might want the robot to pass you the salt and instruct it as follows: "Could you pass me the salt? It is to the left of the stove". Here, the salt is the located object (LO),

J. van den Herik and J. Filipe (Eds.): ICAART 2016, LNAI 10162, pp. 283–301, 2017.
DOI: 10.1007/978-3-319-53354-4_16

because it should be located relative to the reference object (RO, the stove). To find the salt, the robot should interpret this sentence the way you intended it. In the interaction with artificial systems, humans often instruct artificial systems to interact with objects in their environment. To this end, artificial systems must interpret spatial language, i.e., language that describes the locations of the objects of interest. To make the interaction as natural as possible, artificial systems should understand spatial language the way humans do. The implementation of psychologically validated computational models of spatial language into artificial systems might thus prove useful. With these kind of models, artificial systems could begin to interpret and generate human-like spatial language.

[30] were the first to outline a computational framework of the processes that are assumed to take place when humans understand spatial language. Their framework consists of "four different kinds of processes: spatial indexing, reference frame adjustment, spatial template alignment, and computing goodness of fit" [30, p. 500].

Spatial indexing is required to bind the perceptual representations of the RO and the LO to their corresponding conceptual representations. According to [30, p. 499], "the viewer's attention should move from the reference object to the located object". *Reference frame adjustment* consists of imposing a *reference frame* on the RO and setting its parameters (origin, orientation, direction, scale). "The reference frame is a three-dimensional coordinate system [...]" [30, p. 499]. *Spatial template alignment* is the process of imposing a *spatial template* on the RO that is aligned with the reference frame. A spatial template consists of regions of acceptability of a spatial relation. Every spatial relation is assumed to have its own spatial template. Finally, *computing goodness of fit* is the evaluation of the location of the LO in the aligned spatial template.

Trying to identify possible nonlinguistic mechanisms that underlie the rating of spatial prepositions, [35] developed a cognitive model: the Attentional Vector Sum (AVS) model.[1] This model – based on the assumption that goodness-of-fit ratings for spatial prepositions against depicted objects reflect language processing – accounts for a range of empirical findings in spatial language processing. A central mechanism in the AVS model concerns the role of attention for the understanding of spatial relations.

Direction of the Attentional Shift. Previous research has shown that visual attention is needed to process spatial relations ([28–30]; see [7] for a review). The AVS model has formalized the role of visual attention. Although [35] do not explicitly talk about attentional shifts, the AVS model can be interpreted as assuming a shift of attention from the RO to the LO. [35] motivate the implementation of attention based on studies conducted by [28] and [29, p. 115]: "The linguistic distinction between located and reference objects specifies a direction for attention to move – from the reference object to the located object." (See also [30, p. 499]: "the viewer's attention should move from the reference object to

[1] Apart from the AVS model, a range of other computational models of spatial language processing were also proposed, e.g., [5, 16, 19, 36, 39].

the located object".) But are humans actually shifting their attention in this direction while they are understanding a spatial preposition?

Evidence for shifts of covert attention comes from studies in the field of cognitive neuroscience by Franconeri and colleagues [15,38]. Using EEG, [15] showed that humans shift their covert attention when they process spatial relations. In their first experiment, they presented four objects of which two had the same shape but different colors. Two objects were placed to the right and two objects were placed to the left of a fixation cross such that two different shapes appeared on each side of the cross. Participants had to fixate the fixation cross and judge whether, say, the orange circle was left or right of the cyan circle. After the stimulus display was shown, participants chose one spatial relation out of two possible arrangements on a response screen (cyan circle left of orange circle or orange circle left of cyan circle). During the experiment, event-related potentials were recorded. All experiments reported in [15] revealed that participants shifted their attention from one object to the other object, although they had been instructed to attend to both objects simultaneously. However, the role of the *direction* of these shifts remained unclear in [15].

In another experiment, [38] presented questions like "Is red left of green?" to participants. Subsequently, either a red or a green object appeared on the screen, followed shortly afterwards (0–233 ms) by a green or a red object respectively. By manipulating the presentation order of the objects, a shift of attention was cued. Participants were faster to verify the question if the presentation order was the same as the order in the question. [38] interpreted this as evidence that the perceptual representation of a spatial relation follows its linguistic representation.

Evidence that a shift of attention from the RO to the LO as suggested in the AVS model is not necessary for understanding spatial language has been recently reported by [3], who conducted a visual world study. Here, participants inspected a display and listened to spoken utterances while their eye movements were recorded. Note that [3] investigated *overt* attention ([15,38] studied *covert* attention). [3] presented sentences with two German spatial prepositions (*über* [*above*] and *unter* [*below*]) across four different tasks. The RO and the LO of the sentence as well as a competitor object (not mentioned in the sentence) were presented on a computer screen. In their first experiment, participants verified the spatial sentence as quickly as possible, even before the sentence ended. In their second experiment, participants also verified the sentence, but they had to wait until the sentence was over. The third experiment consisted of a passive listening task, i.e., no response was required from the participants. Finally, in the fourth experiment, a gaze-contingent trigger was used: the competitor object and either the LO or the RO were removed from the display after participants had inspected the LO at least once.

The results from this study revealed that participants shifted their overt attention from the RO to the LO, as predicted by the AVS model. However, the task modulated the presence of these shifts. These shifts were only frequent in the post-sentence verification experiment (experiment 2), but infrequent in the other experiments. Crucially, if participants did not shift their attention from the RO to the LO, they performed equally well (as accuracy was not affected) – i.e., they were able to understand the sentence without shifting their attention overtly from the RO to the LO.

By contrast, participants frequently shifted gaze overtly from the LO towards the RO (in line with the incremental interpretation of the spoken sentence). This suggested that people may be able to apprehend a spatial relation with an overt attentional shift from the LO to the RO (and not from the RO to the LO as suggested by the AVS model).

Thus, the direction of the attentional shift as implemented in the AVS model conflicts with recent empirical findings. We propose a modified version of the AVS model: the reversed AVS (rAVS) model, for which the attentional shift has been reversed. Instead of a shift from the RO to the LO, we implemented a shift from the LO to the RO. We designed the rAVS model otherwise to be as similar as possible to the AVS model. By doing so, we can isolate the influence of the reversed shift on the performance of the two models.

2 The Models

In this section, we first describe the AVS model, since the proposed rAVS model is based on the structure of the AVS model and modifies some parts of it. Next, we introduce the rAVS model.

2.1 The AVS Model

[35] proposed a cognitive model of spatial term comprehension: the Attentional Vector Sum (AVS) model. The AVS model takes the 2D-location and the 2D-shape of a RO, the 2D-location of a LO, and a spatial preposition as input and computes an acceptability rating (i.e., how well the preposition describes the location of the LO relative to the RO). In the following, we are presenting how the AVS model processes the spatial relation between the RO and the LO and how it computes the acceptability rating. The AVS model consists of two components: The angular component and the height component. Figures 1a–c depict the angular component which we describe first. Figure 1d visualizes the height component that we describe thereafter.

Angular Component. First, the AVS model defines the focus F of a distribution of visual attention as the point on top of the RO "that is vertically aligned with the trajector [LO] or closest to being so aligned"[2] [35, p. 277]. Next, the model defines the distribution of attention on every point i of the RO as follows (see Fig. 1a for visualization):

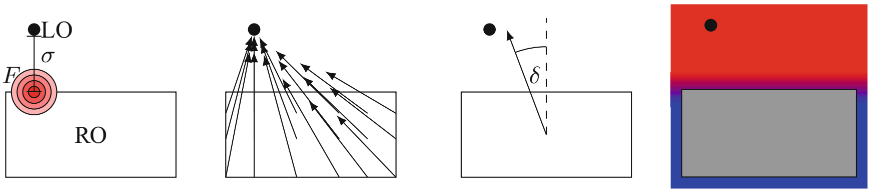
$$a_i = \exp\left(\frac{-d_i}{\lambda \cdot \sigma}\right) \tag{1}$$

[2] In the case of other prepositions, the corresponding part of the RO is chosen for the location of the focus (e.g., the focus lies on the bottom of the RO for *below*).

Here, d_i is the euclidean distance between point i of the RO and the attentional focus F, σ is the euclidean distance between the attentional focus F and the LO, and λ is a free parameter. The resulting distribution of attention is highest at the focal point F and declines exponentially with greater distance from F (see Fig. 1a). Furthermore, the distance σ of the LO to the RO as well as the free parameter λ affect the width of the attentional distribution: A close LO results in a more focused attentional distribution (a large decline of attention from point F) whereas a distant LO results in a more broad attentional distribution (a small decline of attention from point F).

In the next step, vectors v_i are rooted at every point i of the RO. All vectors are pointing to the LO and are weighted with the amount of attention a_i that was previously defined (see Fig. 1b). All these vectors are summed up to obtain a final vector:

$$\overrightarrow{direction} = \sum_{i \in RO} a_i \cdot \vec{v}_i \tag{2}$$



(a) Attentional Distribution. Darker color means higher amount of attention. Attention is maximal at the focal point F. σ denotes the distance of the LO to point F.

(b) Population of vectors weighted with attention as shown in Figure 1a. On every point of the RO one vector is rooted. Every vector points to the LO.

(c) Deviation δ of the final vector from canonical upright (dashed line). The final vector does not necessarily point to the LO, because it is the weighted sum of all vectors in Figure 1b.

(d) Height Component. Red means a value of 1 and blue means a value of 0. The height component is multiplied with the outcome of the angular component.

Fig. 1. Schematized steps of the AVS model developed by [35]. (Color figure online)

The deviation δ of this final vector to canonical upright (in the case of *above*) is measured (see Fig. 1c) and used to obtain a rating with the help of the linear function g(δ) that maps high deviations to low ratings and low deviations to high ratings:

$$\mathrm{g}\,(\delta) = slope \cdot \delta + intercept \tag{3}$$

Both, *slope* and *intercept*, are free parameters and δ is the angle between the sum of the vectors and canonical upright (in the case of *above*):

$$\delta = \angle(\overrightarrow{direction}, upright) \tag{4}$$

Height Component. $\mathrm{g}(\delta)$ is the last step of the angular component. This value is then multiplied with the height component. The height component modulates the final outcome with respect to the elevation of the LO relative to the top of the RO: A height component of 0 results in a low rating, whereas a height component of 1 does not change the output of the angular component. The height component is defined as follows:

$$\mathrm{height}(y_{LO}) = \frac{\mathrm{sig}(y_{LO} - hightop, highgain) + \mathrm{sig}(y_{LO} - lowtop, 1)}{2} \tag{5}$$

Here, *highgain* is a free parameter, *hightop* (or *lowtop*) is the y-coordinate of the highest (or lowest) point on top of the RO, and the $\mathrm{sig}(\cdot,\cdot)$ function is defined as:

$$\mathrm{sig}(x, gain) = \frac{1}{1 + \exp\left(gain \cdot (-x)\right)} \tag{6}$$

The AVS model has four free parameters in total: $\lambda, slope, intercept$, *highgain*. Taken together, the final acceptability rating is computed by the AVS model with the following formula:

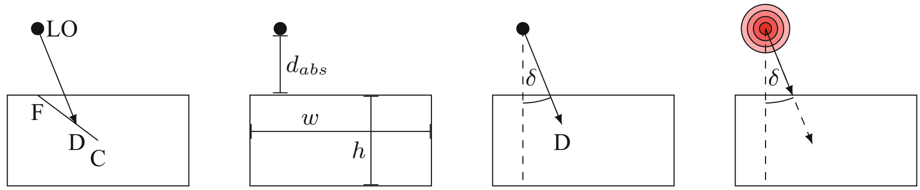
$$\mathrm{above}(LO, RO) = \mathrm{g}\left(\delta\right) \cdot \mathrm{height}(y_{LO}) \tag{7}$$

2.2 The rAVS Model

Although [35] do not explicitly mention shifts of attention, the AVS model can be interpreted as assuming a shift of attention from the RO to the LO: This shift is implemented by the location of the attentional focus and in particular by the direction of the vectors (see Figs. 1a–c). As discussed before, this direction of the attentional shift conflicts with recent empirical findings [3,15,38]. This is why our modified version of the AVS model, the reversed AVS (rAVS) model, implements a shift from the LO to the RO.

To this end, the rAVS model reverses the direction of the vectors in the vector sum in the following way: Instead of pointing from every point in the RO to the LO, the vectors are pointing from every point in the LO to the RO. Since the LO is simplified as a single point in the AVS model, the vector sum in the rAVS model consists of only one vector. The end point of this vector, however, must be defined, since the RO has a mass.

In the rAVS model, the vector end point D lies on the line between the center-of-mass C of the RO and the proximal point F (see Fig. 2a). Here, F is the same point as the attentional focus in the AVS model. Depending on the relative distance of the LO, the vector end point D is closer to C (for distant LOs) or closer to F (for close LOs). Thus, the center-of-mass orientation is more important for distant LOs, whereas the proximal orientation becomes important for close LOs, which corresponds to the rating pattern found by [35, experiment 7]. The width of the attentional distribution in the AVS model has a similar effect.



(a) Vector destination D lies always on the line that connects the center-of-mass of the RO (C) and F.

(b) Relative distance: Absolute distance d_{abs} divided by the width w and height h of the RO.

(c) Deviation δ of the vector $\overrightarrow{LO, D}$ from canonical downwards (dashed line).

(d) The attentional distribution does not change the deviation δ.

Fig. 2. Schematized steps of the rAVS model.

In the rAVS model, the distance of a LO is considered in relative terms, i.e., the width and height of the RO change the relative distance of a LO, even if the absolute distance remains the same (see Fig. 2b). The relative distance is computed as follows:

$$d_{rel.}(LO, RO) = \frac{|LO, P|_x}{RO_{width}} + \frac{|LO, P|_y}{RO_{height}} \tag{8}$$

Here, P is the proximal point in the intuitive sense: The point on the RO that has the smallest absolute distance to the LO. F is guaranteed to lie on top of the RO, whereas P can also be at the left, right, or bottom of the RO. If P is on top of the RO, P equals F.

Furthermore, the computation of the vector end point D is guided with an additional free parameter α (with $\alpha \geq 0$). The new parameter α and the relative distance interact within the following linear function to obtain the new vector destination D:

$$D = \begin{cases} \overrightarrow{LO, C} + (-\alpha \cdot d_{rel.} + 1) \cdot \overrightarrow{CF} & \text{if } (-\alpha \cdot d_{rel.} + 1) > 0 \\ C & \text{else} \end{cases} \tag{9}$$

The direction of the vector $\overrightarrow{LO, D}$ is finally compared to canonical downwards instead of canonical upright (in the case of *above*, see Fig. 2c) – similar to the angular component of the AVS model:

$$\delta = \angle(\overrightarrow{LO, D},\ \ downwards) \tag{10}$$

As in the AVS model, this angular deviation is then used as input for the linear function $g(\delta)$ (see Eq. 3) to obtain a value for the angular component. Note that a comparison to downwards is modeled, although the preposition is *above*. [38, p. 7] also mention this "counterintuitive, but certainly not computationally difficult" flip of the reference direction in their account.

In the rAVS model, the attentional focus lies on the LO. In fact, however, the location of the attentional focus as well as the attentional distribution do not matter for the rAVS model, because its weighted vector sum consists of only one single vector (due to the simplified LO). Since the length of the vector sum is not considered in the computation of the angle (neither in the AVS[3] nor in the rAVS model), the amount of attention at the vector root is not of any importance for the final rating (as long as it is greater than zero, see Fig. 2d).[4]

The height component of the AVS model is not changed in the rAVS model. So, it still takes the y-value of the LO as input and computes the height according to the grazing line of the RO (see Eq. 5). As in the AVS model the final rating is obtained by multiplying the height component with the angular component (see Eq. 7).

3 Model Comparison

In the previous section, we have presented the AVS model by [35] and proposed the rAVS model, since the AVS model conflicts with recent empirical findings regarding the direction of the attentional shift [3,15,38]. But how does the rAVS model perform in comparison to the AVS model?

[35] conducted seven acceptability rating experiments and showed that the AVS model was able to account for all empirical data from these experiments. These data consist of acceptability ratings for $n = 337$ locations of the LO above 10 different types of ROs. We evaluated the rAVS model on the same data set[5] to assess its performance using three different model simulation methods: Goodness-Of-Fit (GOF, Sect. 3.1), Simple Hold-Out (SHO, Sect. 3.2), and Model Flexibility Analysis (MFA, Sect. 3.3). We introduce each of these simulation methods before we present its results.

Both models and all simulation methods were implemented in `C++` with the help of the `Computational Geometry Algorithms Library` [11]. The `C++` source code is available under an open source license from [21]. For all simulations, we constrained the range of the model parameters in the following way:

$$\frac{-1}{45} \leq slope \leq 0 \tag{11}$$

$$0.7 \leq intercept \leq 1.3 \tag{12}$$

$$0 \leq highgain \leq 10 \tag{13}$$

$$0 < \lambda \leq 5 \tag{14}$$

$$0 < \alpha \leq 5 \tag{15}$$

[3] [35, p. 276]: "A central feature of this [angular] characterization of spatial term acceptability is that it is dependent only on the direction, not the length, of the vector connecting the landmark [RO] to the trajector [LO]."

[4] Therefore, the rAVS model does not need to compute a vector *sum* nor does it rely on an underlying attentional distribution and thus has a lower computational complexity. This lower computational complexity, however, originates from the simplification of the LO. Accordingly, these considerations are also only valid for simplified LOs.

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

3.1 Goodness-Of-Fit (GOF)

Method. The Goodness-Of-Fit (GOF) measures how well a model fits given data. We fitted both models to the n rating data points from [35] by minimizing the normalized Root Mean Square Error (nRMSE):

$$nRMSE = \frac{\sqrt{\frac{1}{n}\sum_i^n (data_i - modelOutput_i)^2}}{rating_{max} - rating_{min}} \tag{16}$$

To this end, we used a method known as simulated annealing, a variant of the Metropolis algorithm [31]. This method estimates the free parameters of the model in order to minimize the nRMSE and has the advantage to not get stuck in local minima. The nRMSE gives us a Goodness-Of-Fit (GOF) value. In contrast to the non-normalized RMSE, the normalized RMSE can be compared throughout rating experiments with different rating scales, because it always has a range from 0 to 1: An nRMSE of 0.0 means best performance (the model is able to exactly reproduce the empirical data), an nRMSE of 1.0 means worst performance (model output and data are maximally different).

Results. Figure 3 shows the GOF results for fitting both models to all data from [35]. The model parameters for the plotted GOFs can be found in Table 1. First of all, both models are able to account for the data very closely as is evident from the overall low nRMSE (<0.08 for both models). The rAVS model has a slightly worse GOF value than the AVS model but the difference to the GOF value of the AVS model is very low (difference < 0.005) which renders this difference inconclusive. Also, the GOF values change slightly with each new estimation due to the random nature of the parameter estimation method. The most important conclusion one can draw from the GOF values is whether the models are able to fit the data at all and this is the case for the models and the data under consideration. Assessing the relative performance of more than one model solely with their GOF, however, should be done very carefully.

[37] provide a thorough discussion of the theoretical problems of using GOF as the only measure of model performance. Related to the problems discussed by [37], [34] focus on the specific problem of *overfitting*: A more flexible model might obtain a better GOF just because it is more flexible. *Model flexibility*[6] is the ability of a model to produce arbitrary output: The more different the possible output of a model, the more flexible the model. A more flexible model might fit the noise in the given data better than a less flexible model. Although this results in a better GOF, it does not add anything to the explanatory power of the model in question. We tried to overcome the problem of overfitting by applying the Simple Hold-Out (SHO) method as outlined in the next section.

[6] Following [41] we favor the term *model flexibility* over *model complexity* (used by, e.g., [34]). Both terms mean the same.

3.2 Simple Hold-Out (SHO)

Method. To control for the problem of overfitting, we applied a cross-validation method that takes model flexibility into account: the Simple Hold-Out (SHO) method described in [40]. [40] showed that this method performs very well in comparison to other model comparison methods. In the SHO method, the data set is split into a training and a test set. Model parameters are estimated on the training set and used to compute an nRMSE on the test set. This nRMSE is also called prediction error, because it is the error the model makes for predicting "unseen" data (the test set).[7] This procedure is repeated several times with different, random splits of the data. The median of the prediction errors is the final outcome of the SHO method.

The results presented here were computed with 101 iterations of the SHO method. In each iteration 70% of the data was used as training data and 30 % was used as test data. Moreover, we computed 95% confidence intervals of the SHO median by using 100,000 bootstrap samples with the help of the `boot` package for `R` [6].

Results. The SHO results are plotted next to the GOF results in Fig. 3. These results reveal that the slightly better GOF of the AVS model might be the result of a light overfitting, because both models obtain similar SHO values with overlapping confidence intervals for the SHO values. That is, both models perform equally well and cannot be distinguished on these data using the SHO method.

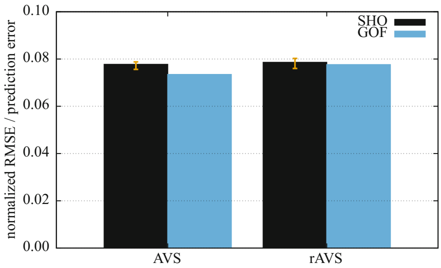


Fig. 3. GOF and SHO results for the AVS and the rAVS model for fitting all data from [35]. Error bars show 95% confidence intervals computed with 100,000 bootstrap samples.

Table 1. Values of the model parameters to achieve the GOFs shown in Fig. 3. The λ parameter of the rAVS model does not change the output of the rAVS model, see Fig. 2d.

	AVS	rAVS
slope	-0.005	-0.004
intercept	0.973	0.943
highgain	0.083	7.497
λ	0.189	(1.221)
α	–	0.322
nRMSE	0.0735	0.0776

[7] The prediction error is also a measure of *model generalizability*, the property of a model to account for new empirical data, see [34].

3.3 Model Flexibility Analysis (MFA)

Method. The Model. Flexibility Analysis (MFA) proposed by [41] tries to account for another problem of using GOFs as a criterion for model evaluation stated by [37]: A model that achieves a good fit to empirical data might also fit other (possibly non-empirical) data very well. Put differently, the fact that a model fits data well does not exclude the possibility that the model might predict outcomes that humans would never generate. If the model also fits non-empirical data[8], its good fit to empirical data becomes less impressive.

Note that this problem is different from the problem for which we applied the SHO method. The SHO method accounts for overfitting, i.e., if a model fits too much noise, it will obtain a good GOF, but a worse SHO result (because predicting unseen data does not benefit from a closer fit to noisy data). Although the SHO method operates with predicting unseen data, it cannot be used for claims about all possible model predictions. This is because the SHO method uses (random) subsets of the same data set. All these subsets are in the same region of the space of possible data, namely, the region of empirically observed data. That is, a model can obtain good SHO values although it has a high flexibility (i.e., although it can generate a great range of different, possibly non-empirical data).

The MFA, however, was explicitly designed to quantify the size of the space of all possible outcomes a model can generate – regardless of whether these outcomes were empirically observed or not. This information can then be used as a measure to “know how impressed to be that theory and observation are consistent” [37, p. 359]. A good fit of a model is more impressive if the model generates (almost) only empirically observed data. A good fit is less impressive if the model also generates a great range of non-empirical data.

Given input to a model (in our case ROs and LOs), the MFA computes all possible model outputs (in our case acceptability ratings) by enumerating the whole space of the free parameters of the model. The outcome of the MFA is the proportion ϕ of these outputs to the size of the data space, where the data space contains all hypothetical possible data:

$$\phi = \frac{\text{number of different model outputs}}{\text{number of all possible data points}} \tag{17}$$

If ϕ is low, the model is only able to compute (and thus fit) a small proportion of theoretically possible data. The lower ϕ, the more strongly the model constrains its possible outcomes, i.e., the less flexible is the model. If ϕ is high, the model can compute a great range of possible data. With a high ϕ the model only weakly constrains its output, i.e., the model is highly flexible. As an example consider one RO with two LOs and a rating scale from 0 to 9. The space

[8] Note that it is difficult to call data “non-empirical”. You can tell what people do, but it is harder to tell what people do not do. Given the right study design, previously considered “non-empirical” data might become empirical. Nevertheless, the greater the range of model predictions, the higher the probability that some of these predictions are at least implausible (or conflict with other generated predictions).

of theoretically possible data is then two-dimensional (two ratings) and each dimension ranges from 0 to 9. Thus, there are $10 \cdot 10 = 100$ theoretically possible data points, if we only consider integer ratings. If a model is able to compute 20 of these rating patterns, it will get the proportion $\phi = \frac{20}{100} = 0.2$, a model that can compute 80 rating patterns results in $\phi = 0.8$.

We computed the MFA proportions on the stimuli that [35] used. The dimension of the data space and each model output is 337, because [35] used 337 locations of the LO in total. We split the range of each of the four free parameters of both models into 50 intervals and computed the model output for each of these 50^4 parameter sets (with a rating scale from 0 to 1). This gave us all outputs the models can generate with these parameter sets. Then, we divided the number of different model outputs by the number of all possible data points (see Eq. 17). To determine whether two model outputs are equal, the MFA uses a grid over the data space. If two model outputs are in the same cell of the grid, they are considered equal. [41] suggest to split each dimension of the data space into $\sqrt[n]{j^k}$ cells, where n is the number of dimensions of the data space (in our case 337), k is the number of free parameters (in our case 4) and j is the number of intervals for each parameter (in our case 50). Accordingly, for our simulations each dimension of the data space should be split into 1.047528 cells. Since we used a rating scale from 0 to 1 for the computation of the MFA, this means that every rating between 0 and 0.954628 is mapped to the first cell and every rating between 0.954628 and 1 is mapped to the second cell. That is, almost all ratings are considered to be equal. This might not be be meaningful in our context, thus, we also considered a different splitting of the data space: Since [35] used a rating scale from 0 to 9 resulting in 10 possible, distinct ratings for each LO, we also computed the MFA with 10 cells for each dimension of the data space without changing the number of the model predictions.[9]

Results. The results of the MFA are shown in Table 2. The worst possible ϕ value is 1.0 (a model that generates all possible data), the lowest possible ϕ value is 0.0 (a model that generates no data). Since all ϕ values in Table 2 are relatively low, both models have a relatively low flexibility. However, regardless of the number of cells in the potential data space, the AVS model obtains higher ϕ values than the rAVS model. Thus, the AVS model is more flexible than the rAVS model which makes the good performance of the AVS model less impressive than the good performance of the rAVS model.

[9] One could argue that we did not compute enough model predictions to define 10 cells on each dimension of the data space. However, if we want to follow the suggestion from [41] and use $\sqrt[n]{j^k}$ cells, we would need more model predictions due to the high-dimensional data space ($n = 337$), namely $j^k = 10^{337}$, i.e. $j = 10^{\frac{337}{4}}$. Unfortunately, splitting each parameter range into $j = 100$ instead of $j = 50$ intervals already resulted in an unmanageable amount of data.

Table 2. Results of the Model Flexibility Analysis (MFA). The lower the ϕ value, the less flexible the model.

Number of cells for each dimension of the data space	AVS	rAVS
$\sqrt[n]{j^k} = 1.047528$	$\phi = 0.00041952$	$\phi = 0.00029248$
10	$\phi = 3.22139 \times 10^{-332}$	$\phi = 2.7116 \times 10^{-332}$

3.4 Discussion

With the GOF and SHO simulations, we showed that both models are able to account equally well for the data from [35] and that this good performance is not the result of overfitting. Considering only these results, both attentional shifts are equally well supported. However, the AVS model also showed a greater flexibility in the MFA: the AVS model computes a greater range of possible outputs than the rAVS model. The lower flexibility of the rAVS model thus makes the good performance of the rAVS model more impressive than the good performance of the AVS model. This is because the rAVS model achieves the same performance while it also more strongly constrains the space of the model output. The rAVS model thus can be more easily falsified than the AVS model.

We asked where the greater flexibility of the AVS model originates. Computationally, the biggest difference of the two models is the computation of the final vector: The AVS model uses a weighted vector sum whereas the rAVS model only uses a linear function.[10] Thus, the vector sum seems to be the main source of the greater flexibility of the AVS model. Our results suggest that the vector sum is more flexible than is needed for the empirical data. This does not necessarily mean that the shift from the RO to the LO as assumed in the AVS model is less supported than the reversed shift. Rather, the way the shift is implemented in the AVS model might be more complex than needed.[11] However, [35] motivated the implementation of the vector sum as a possible non-linguistic mechanism underlying the linguistic process. More specifically, they used the vector sum, because it seems to be a widely used representation of direction in the brain (see discussion below). Accordingly, the vector sum is a crucial part of the model. Our results show that the AVS model has an overall low flexibility but they also show that the weighted vector sum is more flexible than the linear function used by the rAVS model. However, we did not present a model that performs better, just a model that performs equally well – with a lower flexibility. Moreover, the rAVS model does not offer a competing explanation of how the vector is

[10] Conceptually, the rAVS model also uses a vector sum on the LO. However, since the LO is simplified as a single point for the current model input, the rAVS model in fact does not compute a vector sum.

[11] A model that implements a shift from the RO to the LO without using a vector sum could be a modified rAVS model: One could change the direction of the vector and the reference direction. Computationally, this model computes the same output as the rAVS model.

computed (mainly because its motivation was different). At the neuronal level, this computation could still be done with a vector sum but the rAVS model does not provide details about the neuronal level.

Although our model simulations do not result in the support of one of the two shifts in question, they raise the question to which degree the attentional shift from the RO to the LO as proposed by [29] and [30] is the only shift that is implicated in the processing of spatial relations. The results from [3] suggest that humans perform both shifts, but that an overt shift from the LO to the RO alone (as in the rAVS model) can be enough to apprehend the spatial relation between the objects. The shift back (from the RO to the LO) could be a way to double-check the goodness-of-fit of the spatial preposition. Our results support this by showing that the rAVS model – that assumes only the shift from the LO to the RO – can account for the data from [35].

4 Conclusion

We proposed a new cognitive model for spatial language understanding: the rAVS model. This model is based on the AVS model by [35] but integrates recent psycholinguistic and neuroscientific findings [3,15,38] that conflict with the assumption of the direction of the attentional shift in the AVS model. In the AVS model, attention shifts from the RO to the LO; in the rAVS model, attention shifts from the LO to the RO. We assessed both models using the data from [35] and found that both models perform equally well, while the rAVS model is less flexible than the AVS model. Accordingly, our model simulations favor the rAVS model. Since the lower flexibility of the rAVS model originates from the lack of using a vector sum, however, the advantage of the rAVS model does not result in the favor of any of the two directionalities of the attentional shift. However, we showed that both directionalities can account for the empirical data.

Theoretical Contribution. [35] developed the AVS model with the goal to identify possible nonlinguistic mechanisms that underlie spatial term rating. To this end, they implemented two independent observations in the AVS model: First, the importance of attention to understand spatial relations and second, the neuronal representation of a motor movement as a vector sum. So, the main goal of the AVS model was not to examine the direction of the shift of attention but rather to describe linguistic processes with nonlinguistic mechanisms.

Although the focus of the AVS model was not on the direction of the attentional shift, the model implies a shift from the RO to the LO. [35] motivated the use of a vector sum because it seems to be a widely used representation of direction in the brain. [17] found that the direction of an arm movement of a rhesus monkey can be predicted by a vector sum of orientation tuned neurons. [27] provide evidence for a similar representation for saccadic eye movements. Eye movements (overt attention) are motor movements that are closely connected to covert visual attention: “Many studies have investigated the interaction of overt and covert attention, and the order in which they are deployed. The consensus is that covert attention precedes eye movements [...].” [10, p. 1487] Although the authors of the AVS model do not explicitly speak about which movement the vector sum in their model represents nor clearly specify the kind of attention in the model, it seems reasonable to interpret the direction of the vector sum in the AVS model as the direction of a shift of attention that goes from the RO to the LO.

Our aim was to implement the most recent findings of attentional mechanisms into the AVS model. To this end, we designed the rAVS model as similar as possible to the AVS model. So, the rAVS model follows the same basic concepts while it integrates the most recent findings. We do not claim that the nonlinguistic mechanisms proposed in the AVS model do not happen – rather, we propose an alternative way of how they might take place. That is, on the neuronal level, the orientation of the vector in the rAVS model that points from the LO to the RO could still be computed by a weighted population of neurons (similar to the attentional vector sum) but the rAVS model does not provide such details. Its focus lies on a more abstract level that concerns the *direction* of the attentional shift and not the detailed computation of this shift. Keeping the same basic concepts as the AVS model, the rAVS model accounts for the same data equally well – and also for the recent empirical findings regarding the direction of the attentional shift.

4.1 Future Work

Modeling Both Shifts. The success of both the rAVS model and the AVS model support the existence of *both* directionalities of the attentional shift. It might well be that people shift their attention in both directions during the processing of spatial relations – depending on the task and the linguistic input. Accordingly, a model that implements both attentional shifts might fit more data than the AVS or the rAVS model alone.[12]

It might be interesting to investigate this possibility by creating a model that allows both shifts of attention. Such a model should be applicable to more types of experimental data than the AVS model and the rAVS model (which both can only account for acceptability rating data). In particular, the model with both shifts should also specify when in time what type of attentional shift occurs and how long the computation takes. This model could then be fitted to a greater range of data, like real-time eye movement data from visual world studies (e.g., [3]) or reaction time data (e.g., [38]). Modeling different tasks would give more insight into the role of the attentional shift.

Modeling the LO. The main reason for the lower flexibility as well as the lower computational complexity of the rAVS model is the simplification of the LO as a single point. There is evidence, however, that geometric features of the LO also affect acceptability ratings [1,2,4]. A comprehensive model of spatial language thus should also model the LO in more detail. Accordingly, we are planning to extend the representation of the LO in the rAVS model by giving a mass to it. This would give us the opportunity to see first how the rAVS model

[12] We thank an anonymous reviewer for suggesting this idea.

deals with the situation where the computation of a vector sum is necessary to determine the angular deviation. Second, an extended LO might affect the role of the attentional distribution in the rAVS model.

We are also interested in modifying the use of the height component in the rAVS model. At the moment, the rAVS model applies the same computation as the AVS model for the height component: the y-coordinate of the LO is compared relative to the top of the RO (see Fig. 1d). In the rAVS model, the attentional focus is located on the LO. So, it would be more consistent if the location of the LO were taken as the baseline for the comparison with the location of the RO. Thus, we want to reverse the computation of the height component such that the grazing line lies on the bottom of the LO.

Model Distinction. To tease apart the two models and evaluate the accuracy of their predictions, we are currently analyzing the models with two more model simulation methods: the *landscaping* analysis proposed by [32,42] and an algorithm called *Parameter Space Partitioning* (PSP) proposed by [20,33].

Landscaping provides an overview how data and models behave to each other and how informative a specific data set is in distinguishing two models. 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) and then both models fit these data. Landscaping provides a measure of what is called *model mimicry* by [42]: The ability of a model to account for data generated by another model. Each model should fit the 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. We are currently applying the landscaping method on the stimuli from [35]. On a different set of stimuli, a landscaping analysis confirmed the greater flexibility of the AVS model (i.e., the AVS model mimics the rAVS model but not vice versa, see [24,25]).

The PSP algorithm is a Markov chain Monte Carlo (MCMC) based method and searches in the parameter space of the models for regions of patterns that are qualitatively different. First results confirm the high flexibility of the AVS model (i.e., the AVS model is able to generate many patterns that are qualitatively different by using different sets of parameters). The rAVS model, however, generates fewer patterns with a qualitative difference (see [22] for details). To test the predictions revealed by the PSP analysis we conducted an empirical rating study with the same stimuli. More on the results of this study can be found in [24,25].

Functionality. The AVS model does not account for any effects of the functionality of objects on spatial language comprehension, although there is evidence that – beside purely geometric effects – functional interactions between objects also affect the use of spatial prepositions [8,9,12–14,18].

For instance, [9] conducted an object placement task, where participants had to place a toothpaste tube above a toothbrush. They showed that the toothpaste tube was not placed above the center-of-mass of the toothbrush, but rather above the bristles of the toothbrush – that is, at the location where both objects can functionally interact. Objects with a smaller amount of functional interaction (here, a tube of oil paint) were placed more above the center-of-mass of the toothbrush instead of above the bristles.

Despite this evidence, the AVS model (and thus also our rAVS model) only considers geometric representations of the RO and the LO. For the AVS model, however, a range of extensions that integrate functionality were already proposed [8,26]. Since the rAVS model is designed to be as similar as possible to the AVS model, these functional extensions might also be applicable for the rAVS model.

Implementing the Models in Artificial Systems. In order to implement these models into artificial systems, additional steps are necessary. The models were designed to model spatial language *understanding*. The models thus produce an acceptability rating given a RO, a LO, and a preposition. As part of an artificial system that *interprets* spatial language, the models can be used straightforwardly: Given a spatial utterance and a visual scene, the models can be used to compute acceptability ratings for all points around the RO (i.e., a spatial template). The artificial system then starts the search for the LO at the point with the highest rating. To *generate* spatial language with the help of these models, one could imagine the following steps: Compute the acceptability ratings of different spatial prepositions (e.g., above, below, to the left of, in front of, ...) and subsequently pick the one with the highest rating.

Simulating the models showed that the computation of the attentional vector sum is computationally more expensive than the linear function that is used by the rAVS model. Thus, the rAVS model provides a shortcut for the computation of the final vector. In particular, this is interesting for the implementation of the model in real-time robotic systems that often have constrained computational resources.

In conclusion, we proposed a modified version of the AVS model: the rAVS model. The rAVS model accounts for the same empirical data as the AVS model while integrating additional recent findings regarding the direction of the attentional shift that conflict with the assumptions of the AVS model.

Acknowledgments. 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 authors would also like to thank two anonymous reviewers for their useful comments and suggestions.

References

1. Burigo, M.: On the role of informativeness in spatial language comprehension. Ph.D. thesis, School of Psychology, University of Plymouth (2008)
2. Burigo, M., Coventry, K.R., Cangelosi, A., Lynott, D.: Spatial language and converseness. Q. J. Exp. Psychol. **69**(12), 2319–2337 (2016). doi:10.1080/17470218.2015.11248942016
3. Burigo, M., Knoeferle, P.: Visual attention during spatial language comprehension. PLoS ONE **10**(1), e0115758 (2015)

4. Burigo, M., Sacchi, S.: Object orientation affects spatial language comprehension. Cogn. Sci. **37**(8), 1471–1492 (2013)
5. Cangelosi, A., Coventry, K.R., Rajapakse, R., Joyce, D., Bacon, A., Richards, L., Newstead, S.N.: Grounding language in perception: a connectionist model of spatial terms and vague quantifiers. Prog. Neural Process. **16**, 47 (2005)
6. Canty, A., Ripley, B.: Boot: Bootstrap R (S-Plus) Functions (2015). R package version 1.3-15
7. Carlson, L.A., Logan, G.D.: Attention and spatial language. In: Itti, L., Rees, G., Tsotsos, J.K. (eds.) Neurobiology of Attention, pp. 330–336. Elsevier, Amsterdam (2005). Chap. 54
8. Carlson, L.A., Regier, T., Lopez, W., Corrigan, B.: Attention unites form and function in spatial language. Spat. Cogn. Comput. **6**(4), 295–308 (2006)
9. Carlson-Radvansky, L.A., Covey, E.S., Lattanzi, K.M.: "What" effects on "where": functional influences on spatial relations. Psychol. Sci. **10**(6), 516–521 (1999)
10. Carrasco, M.: Visual attention: the past 25 years. Vis. Res. **51**(13), 1484–1525 (2011)
11. CGAL: Computational geometry algorithms library. http://www.cgal.org
12. Coventry, K.R., Garrod, S.C.: Saying, Seeing, and Acting: The Psychological Semantics of Spatial Prepositions. Essays in Cognitive Psychology, Psychology Press, Taylor and Francis, Hove and New York (2004)
13. Coventry, K.R., Lynott, D., Cangelosi, A., Monrouxe, L., Joyce, D., Richardson, D.C.: Spatial language, visual attention, and perceptual simulation. Brain Lang. **112**(3), 202–213 (2010)
14. Coventry, K.R., Prat Sala, M., Richards, L.: The interplay between geometry and function in the comprehension of over, under, above, and below. J. Mem. Lang. **44**(3), 376–398 (2001)
15. Franconeri, S.L., Scimeca, J.M., Roth, J.C., Helseth, S.A., Kahn, L.E.: Flexible visual processing of spatial relationships. Cognition **122**(2), 210–227 (2012)
16. Gapp, K.-P.: An empirically validated model for computing spatial relations. In: Wachsmuth, I., Rollinger, C.-R., Brauer, W. (eds.) KI 1995. LNCS, vol. 981, pp. 245–256. Springer, Heidelberg (1995). doi:10.1007/3-540-60343-3_41
17. Georgopoulos, A.P., Schwartz, A.B., Kettner, R.E.: Neuronal population coding of movement direction. Science **233**, 1416–1419 (1986)
18. Hörberg, T.: Influences of form and function on the acceptability of projective prepositions in Swedish. Spat. Cogn. Comput. **8**(3), 193–218 (2008)
19. Kelleher, J.D., Kruijff, G.J.M., Costello, F.J.: Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expressions. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 745–752. Association for Computational Linguistics (2006)
20. Kim, W., Navarro, D.J., Pitt, M.A., Myung, I.J.: An MCMC-based method of comparing connectionist models in cognitive science. Adv. Neural Inf. Process. Syst. **16**, 937–944 (2004)
21. Kluth, T.: A C++ implementation of the reversed Attentional Vector Sum (rAVS) model. Bielefeld University (2016). doi:10.4119/unibi/2900103
22. Kluth, T., Burigo, M., Knoeferle, P.: Investigating the parameter space of cognitive models of spatial language comprehension. In: 5. Interdisziplinärer Workshop Kognitive Systeme, Bochum (2016)
23. Kluth, T., Burigo, M., Knoeferle, P.: Shifts of attention during spatial language comprehension: a computational investigation. In: Proceedings of the 8th International Conference on Agents and Artificial Intelligence, vol. 2, pp. 213–222. SCITEPRESS (2016). doi:10.5220/0005851202130222

24. Kluth, T., Burigo, M., Schultheis, H., Knoeferle, P.: Distinguishing cognitive models of spatial language understanding. In: Proceedings of the International Conference on Cognitive Modeling (2016). Poster presented at the ICCM 2016
25. Kluth, T., Burigo, M., Schultheis, H., Knoeferle, P.: Testing the predictions of cognitive models of spatial language comprehension (in preparation)
26. Kluth, T., Schultheis, H.: Attentional distribution and spatial language. In: Freksa, C., Nebel, B., Hegarty, M., Barkowsky, T. (eds.) Spatial Cognition 2014. LNCS (LNAI), vol. 8684, pp. 76–91. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11215-2_6
27. Lee, C., Rohrer, W.H., Sparks, D.L.: Population coding of saccadic eye movements by neurons in the superior colliculus. Nature **332**, 357–360 (1988)
28. Logan, G.D.: Spatial attention and the apprehension of spatial relations. J. Exp. Psychol. Hum. Percept. Perform. **20**(5), 1015 (1994)
29. Logan, G.D.: Linguistic and conceptual control of visual spatial attention. Cogn. Psychol. **28**(2), 103–174 (1995)
30. Logan, G.D., Sadler, D.D.: A computational analysis of the apprehension of spatial relations. In: Bloom, P., Peterson, M.A., Nadel, L., Garrett, M.F. (eds.) Language and Space, pp. 493–530. The MIT Press, Cambridge (1996). Chap. 13
31. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)
32. Navarro, D.J., Pitt, M.A., Myung, I.J.: Assessing the distinguishability of models and the informativeness of data. Cogn. Psychol. **49**(1), 47–84 (2004)
33. Pitt, M.A., Kim, W., Navarro, D.J., Myung, J.I.: Global model analysis by parameter space partitioning. Psychol. Rev. **113**(1), 57–83 (2006)
34. Pitt, M.A., Myung, I.J.: When a good fit can be bad. Trends Cogn. Sci. **6**(10), 421–425 (2002)
35. Regier, T., Carlson, L.A.: Grounding spatial language in perception: an empirical and computational investigation. J. Exp. Psychol.: Gen. **130**(2), 273–298 (2001)
36. Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., Schöner, G.: Autonomous neural dynamics to test hypotheses in a model of spatial language. In: Bello, P., Guarini, M., Mc-Shane, M., Scassellati, B. (eds.) Proceedings of the 36th Annual Conference of the Cognitive Science Society, pp. 2847–2852. Cognitive Science Society, Austin (2014)
37. Roberts, S., Pashler, H.: How persuasive is a good fit? A comment on theory testing. Psychol. Rev. **107**(2), 358–367 (2000)
38. Roth, J.C., Franconeri, S.L.: Asymmetric coding of categorical spatial relations in both language and vision. Front. Psychol. **3**, Article No. 464 (2012)
39. Schultheis, H., Carlson, L.A.: Mechanisms of reference frame selection in spatial term use: computational and empirical studies. Cogn. Sci. (2015). doi:10.1111/cogs.12327
40. Schultheis, H., Singhaniya, A., Chaplot, D.S.: Comparing model comparison methods. In: Proceedings of the 35th Annual Conference of the Cognitive Science Society, pp. 1294–1299. Cognitive Science Society, Austin (2013)
41. Veksler, V.D., Myers, C.W., Gluck, K.A.: Model flexibility analysis. Psychol. Rev. **122**(4), 755–769 (2015)
42. Wagenmakers, E.J., Ratcliff, R., Gomez, P., Iverson, G.J.: Assessing model mimicry using the parametric bootstrap. J. Math. Psychol. **48**(1), 28–50 (2004)