# Visual Attention and the Semantics of Space: Evidence for Two Forms of Symbolic Control

**Matthias Scheutz (mscheutz@nd.edu) and Bradley Gibson (bgibson@nd.edu)**
Departments of Computer Science and Engineering, and Psychology
University of Notre Dame
Notre Dame, IN 46556, USA

## Abstract

In this paper, we investigate the functional differences between word cues and arrow cues in a spatial cuing task and provide a novel computational model fit to the empirical data that provides (1) a conceptually parsimonious explanation of the observed differences and (2) evidence for the existence of two forms of symbolic attentional control. We briefly discuss the implications of the model for theories of spatial reference frames and attentional control.

## Introduction[1]

Over the past thirty years, researchers have used a variety of directional symbols to elicit covert visual orienting within the context of the spatial cuing paradigm (Posner, Snyder, & Davidson, 1980). However, this research has generally proceeded without explicit regard for the processing constraints that the comprehension of such symbols might place on the orientation of attention. In this paper, we further investigate functional differences between word cues and arrow cues and provide a computational model that implies a novel, conceptually parsimonious explanation of the observed differences, while also providing evidence for the existence of two forms of symbolic attentional control.

## Background

Recently, Gibson and Kingstone (in press) have proposed a new taxonomy of spatial cues that is based on the linguistic distinction between projective and deictic spatial relations (see also Logan, 1995). In their study, displays containing two green circles and two red circles were presented in the four cardinal locations, and observers were instructed to report the color of the cued circle (see Fig. 1). The distinction between projective and deictic spatial relations can be understood by considering how a word cue such as "above" and the corresponding arrow cue each refer to spatial locations. Although both cues refer to the circle that appears in the uppermost location in the display, these two cues refer to this location in two semantically different ways. In the word cue condition, the information provided by the cue states that the target is above the cue. In this situation, knowledge of direction is necessary to find the target. This knowledge is thought to derive from a relatively complex process in which observers impose their frame of reference onto the cue and then identify the appropriate pole ("above") of the appropriate axis (Carlson, 2003; Carlson, West, Taylor, & Herndon, 2004; Logan, 1994, 1995; Logan & Sadler, 1996). In contrast, in the arrow cue condition, the information provided by the cue states that the target is there. In this situation, knowledge of direction is not necessary to find the target. One does not need to know that the target is above the cue; rather, one only needs to know that the target is "there" or in "that" location.

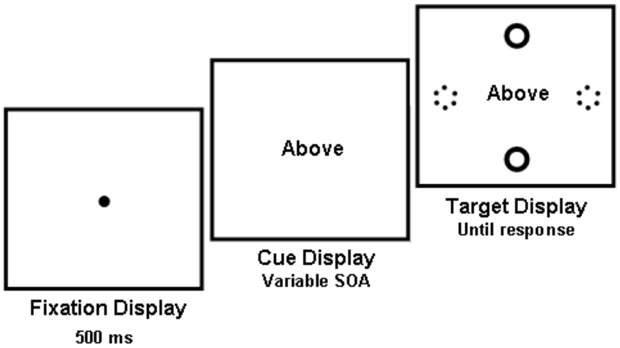

Figure 1: Experimental setup (see text for explanation). Note, solid lines = green; dotted lines = red.

About a decade ago, Logan (1995) proposed an important empirical diagnostic for distinguishing between projective and deictic spatial relations (though he used different terms). After consideration of the computational processes that underlie comprehension of projective terms, Logan proposed the "conceptual frame hypothesis." According to this hypothesis, directions such as "above" and "below" might be easier to access than directions such as "left" and "right." In other words, observers may first need to define the "above" and "below" directions before they can define the "left" and "right" directions. If so, then observers may be able to shift their attention more efficiently in response to above and below cues than they can in response to left and right cues. In contrast to the conceptual frame hypothesis, Logan proposed the "equal accessibility hypothesis" for deictic spatial relations. According to this hypothesis, all four locations should be equally accessible. Consistent with these predictions, Gibson and Kingstone (in press) found that observers were able to shift their attention more efficiently to the "above" and "below" locations than to the "left" and "right" locations when word cues preceded the appearance of the target display by either 0 ms, 250 ms, or 500 ms; but, observers were able to shift their attention to each of these locations equally efficiently when arrow cues were shown. In addition, Gibson and King-

[1] Both authors contributed equally to this paper.

stone also found that color discrimination latencies were overall much slower in the word cue condition than in the arrow cue condition across the three SOAs.

In summary, these previous findings suggest that word cues such as *above*, *below*, *left*, and *right* express projective spatial relations, whereas, arrow cues express deictic spatial relations. We will in the following further investigate this theoretical claim by (1) conducting an additional experiment aimed at elaborating the difference between word cues and arrow cues, and (2) providing a computational connectionist model that is fit to the experimental data. By keeping the proposed components of the model *minimal*, the model will allows to determine a parsimonious set of functional components (necessary to distinguish the processing of word and arrow cues in humans).

## Empirical Experiment and Results

There were two important effects observed in Gibson and Kingstone's (in press) study. First was the cued location effect (RTs in the "above/below" cue condition were faster than RTs in the "left/right" condition) which was obtained exclusively in the word cue condition. The second important result was the cue type effect (RTs in the arrow cue condition were faster than RTs in the word cue condition). Surprisingly, both effects continued to persist even after cue-target SOAs of 500 ms. The cued location effect observed in the word cue condition remained constant across the three SOAs; the cue type effect was found to decrease as a function of SOA, but was not eliminated. Because the time course associated with each of these two effects has important implications for fitting model parameters, the present experiment was designed to provide a more detailed understanding of how these two effects might change over time.

### Method

*Participants.* Thirty-six undergraduates from the University of Notre Dame participated in this experiment in partial fulfillment of a course requirement. Eighteen undergraduates were randomly assigned to either the word cue condition or the arrow cue condition. The observers all reported normal or corrected-to-normal vision.

*Stimuli.* The experimental methodology was based on the experiments reported by Gibson and Kingstone (in press). Three displays were presented on each trial: a fixation display, a cue display, and a target display. The initial fixation display was a small fixation dot ($0.38^o$ in diameter). The cue displays contained one of the two cue types. The word cues were written in capital letters. The words were all $0.68^o$ tall, and ranged in length from $1.18^o$ to $1.94^o$. The word cues replaced the fixation dot when they appeared, as did the arrow cue which subtended $0.48^o$ X $1.18^o$ of visual angle. The target display contained four colored O's that measured $1.26^o$ in diameter and were presented at the four cardinal locations, approximately $4.37^o$ from the central fixation point. Two of the O's were colored red and two were colored green on each trial.

*Procedure and Design.* A typical trial sequence is shown in Fig. 1. Each trial began with a fixation display for 500 ms followed by the cue display. Cue type remained constant within each of the two groups of observers. The cues were presented equally often in each of eight cue-target SOA conditions: 250 ms, 500 ms, 750 ms, 1000 ms, 1250 ms, 1500 ms, 1750 ms, and 2000 ms (the eight SOA conditions were presented randomly during the experimental trials). Cues stayed on throughout the duration of the trial to dissuade observers from using verbal codes to maintain the spatial information conveyed by the cue. The target display then appeared and, together with the cue, remained on the screen until a response was made. The cues were 100% valid and always indicated which one of the four O's was the target; observers' task was to determine as quickly and accurately as possible whether the target O was red or green. The cue referred to each one of the four target locations equally often, and on any given trial, each location was equally likely to contain a red or green O. In this way, observers could not determine (without guessing) how to respond without the aid of the cue These contingencies provided reasonable assurance that observers would process the different cue types equally, even though such processing might differ in complexity. Observers always used their left hand to respond "red" and their right hand to respond "green;" however, for half of the observers the response pad was arrayed horizontally (with "red" to the left of "green") and for the other half the response pad was arrayed vertically (with "red" above "green").

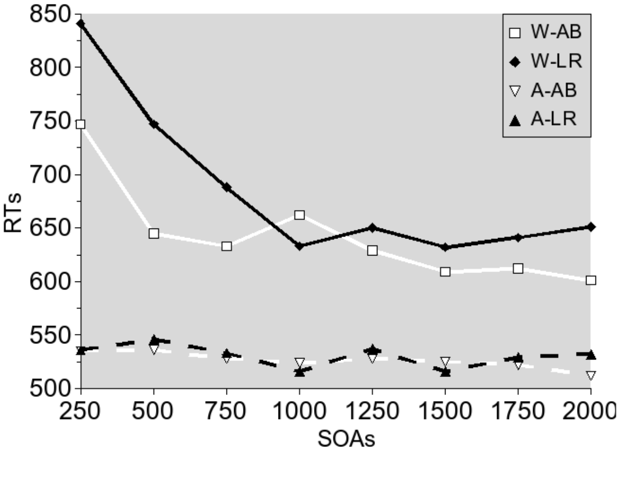

Figure 2: Results from the human experiments.

### Results

Mean correct RTs are shown in Fig. 2 as a function of cue location and SOA in each of the two cue type conditions. The present study was conducted to investigate two issues. The first issue concerned the time course of cued location effect observed between the "above/below" and "left/right" locations in the word cue condition; and, the second issue concerned the time course of the cue type effect observed between the word cue and arrow cue conditions. A three-way mixed analysis of variance (ANOVA) was performed with cue type (word cues vs. arrow cues) as the sole between-subjects variables, and SOA (250 ms vs. 500 ms vs. 750 ms vs. 1000 ms vs. 1250 ms vs. 1500 ms vs. 1750 vs. 2000 ms) and cued location ("above/below" vs. "left/right") as the two within-subjects factors.

As expected, cued location had a significant effect in the word cue condition, but not in the arrow cue condition, $F(1, 34) = 12.52, MSe = 6470.36, p < .01$, for the cued location X cue type interaction. Further ANOVAs conducted for each of the two cue type conditions separately confirmed that there was significant interaction between SOA and cued location in the word cue condition, $F(7, 119) = 2.38, MSe = 6665.42, p < .05$, but there was no evidence that cued location had any effect on performance in the arrow cue condition (all $p's > .15$). Consistent with the findings reported by Gibson and Kingstone (in press), there was a large and relatively consistent cued location effect observed in the 250 ms, 500 ms, and 750 ms SOA conditions (all $p's < .05$); however, the present findings also showed that the cued location effect decreased in the longer SOA conditions. In fact, with the exception of the 1750 ms SOA, which show a relatively small but reliable cued location effect, $F(1, 17) = 4.95, MSe = 1533.68, p < .05$, the effect of cued location was generally found to be small and non-significant when SOA was 1000 ms or greater (all remaining $p's > .05$). Thus, these findings indicate that attention can be shifted more efficiently in response to "above/below" cues than in response to "left/right" cues at relatively short SOAs; however, this advantage is diminished at longer SOAs.

The present results also showed that the overall RT difference observed between the word cue and arrow cue conditions did decrease as SOA increased, $F(7, 238) = 13.17, MSe = 3796.76, p < .001$, for the SOA X cue type interaction. However, the overall RT difference observed between the word cue and arrow cue conditions nevertheless remained significant at each of the eight SOA conditions used in the present study ($p < .05$ or less for each of the eight pair-wise comparisons). Notice also that RTs appeared to have reached asymptotic levels in the both the word cue and arrow cue conditions; thus, the enduring effect of cue type observed in the present study cannot be attributed to the use of insufficiently long SOAs. In summary, the experimental results raise two critical questions:

(Q1) *cued location effect* – why are RTs in the above/below condition faster than in the left/right condition when word cues are shown, but not when arrow cues are shown? Previous answers to this questions have critically involved the notion of a "reference frame" (see the Background Section).

(Q2) *cue type effect* – why are overall RTs in the arrow cue condition faster than the overall RTs in the word cue condition? This is a new effect for SOAs beyond 500 msec for which no detailed hypotheses have been proposed.

## Connectionist Model and Simulations

The purpose of the computational model is to find the simplest architecture that has both psychologically plausible functional components and can be fit to the above experimental data. Consequently, any model of the above task needs to have, at the very least, a component representing the features of the input image, a visual workspace in which visual representations can be processed, an attentional mechanisms that can bias processing in the visual workspace, a conceptual representation of locations and directions, a lexical representation of words, and a decision mechanism to choose a target color.

Given such a model, we will be able to provide explanations for the cued location and type effects witnessed in the present experiments. Specifically, we formulate two hypotheses corresponding to the previous two questions (Q1) and (Q2) that will be tested with the model: (H1) the differences in response times between above/below vs. left/right conditions for words is *solely due* to a difference in connection strength between lexical and semantic representations of above/below vs. left/right; and, (H2) the overall differences in response times between the word cue and arrow cue conditions is due *both* to the direct activation of concepts by arrows (as opposed to the indirect activation of concepts by words via a mediating lexical representation) and the direct activation of processing areas in the visual workspace.

We start with a description of the general model architecture and then proceed to the specification of the particular model used to fit to the human data. We include a brief justification of the employed methodology for parameter fitting and then report the results from simulations with the model.

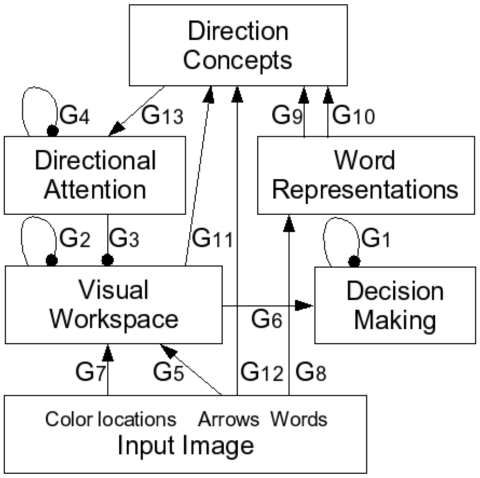

Figure 3: The basic model architecture, consisting of six major architectural components. Lines with arrows and circles depict excitatory or inhibitory connections, respectively (see text for an explanation of the labels of the connections).

## Model Architecture

The general model architecture, depicted in Fig. 3, consists of six main functional divisions or components, each of which comprises several individual computational units that are connected via inhibitory (G1-G4) and excitatory (G5-G13) connections (numbers of units employed in the specific instance of the architecture are given for each component in parentheses):

- The *(pre-processed) input image* (16 units), consisting of representations for colored circles, words, and arrows and their location in the image.
- The *visual workspace* (8 units), which has representations for colored objects and is used for visual computations to determine the target object.
- The *directional attention* (4 units), which can activate or suppress different locations in visual workspace based on attentional focus.
- The *word representations* (4 units), which are activated by word images and, in turn, effect the semantic mapping onto concepts.
- The *direction concepts* (4 units), which represent the direction concepts of "above," "below," "left," and "right."
- The *decision making component* (2 units), which determines the color of the target object via an decision threshold $\theta = 0.2$ (for the difference in activation between the two node representing one of the two colors each).

The employed computational units are simplified versions of the well-known "interactive activation and competition" units (McClelland & Rumelhart, 1988), whose change in activation is given by

$$\Delta act/\Delta t = netin - act \cdot (netin + decay)$$

where $act \in [0, 1]$ is the activation of the unit, $netin \in [0, 1]$ the summed weighted input to the unit and $decay \in [0, 1]$ is a constant decay factor (set to 0.05 for all nodes). Moreover, the sum of all incoming connection weights can be at most 1 (to guarantee that $netin \in [0, 1]$).

## Specific Model Parameters and Parameter Fitting

The specific model we used to model the empirical data from human subjects was intended to be the smallest model that can be fit well. Hence, we included only those computational units that were necessary to complete the task appropriately and excluded other units that would have to be present to implement more accurately the full functionality of a functional component (e.g., we did not include units for word representations in visual workspace as they are not necessary for the explanation of the empirical data).

In the minimal model, units in one component are typically connected to corresponding units in another component (as indicated in Fig. 3). Units in the attention and decision making component are, in addition, fully connected without self-connections via G4 and G1, respectively, and units in the visual workspace representing the two colors "red" and "green" in the four different locations "above", "below", "left", and "right" are connected pairwise via inhibitory links G2. Finally, directional attention units suppress all locations except the one they represent in visual workspace via inhibitory connections G3. The unit representing "left green" circle in the input image, for example, is connected to the unit representing "left green" in the visual workspace via G7, which in turn has inhibitory connections G2 to "left red" (as an object cannot have two colors at the same time), and has excitatory connections G11 to the "left" concept node via G11 (to activate the "left" concept if something is processed in the left area of the visual workspace). The "left" concept node, in turn, has an excitatory connection G13 to the "left" attention node, which suppresses the activation of all other attention nodes via G4 (in a competition process) and, moreover, suppresses the activation of the representations in locations other than "left" in the visual workspace via G3. The "left arrow" node in the center of the input image then corresponds to the prime and activates concepts directly via G12, but also primes the "left" location in visual workspace via G5. The word image "left" in the center of the input image, on the other hand, first activates its lexical representation via G8, which, in turn, activates the direction concept "left" via G10. Note that G10 is used for connections between the words "left" and "right" and their respective direction concepts, while G9 is used for the words "above" and "below" and their respective direction concepts. The reason for separating out these two sets (instead of having one weight group with the same values for all four weights between words and concepts) will become discussed shortly. Note that all connections labeled "G$n$" (i.e., *weight groups*) have the same value for each $n$ (i.e., within a group).

Hence, the model has 13 free parameters that can be used to fit the model to the empirical data. To reduce the number of free parameters, we use the same value (-0.065) for all inhibitory connections (G1 - G4) and require that the remaining excitatory connections be reasonably similar in magnitude (i.e., between 0 and 0.1). Since we wanted to determine the best set of the remaining 9 values such that the difference between the model data and the human data was a low as possible, it was critical to determine a mapping between the response time data from humans and the model simulation. As in previous models (e.g., Scheutz & Eberhard, 2004), we used the simple mapping $f(t) = t/10$ from milliseconds into update cycles (i.e., 100 msec of real-time corresponds to 10 update cycles in the model). Given the mapping $f$, it is then possible to apply "external inputs" (e.g., the prime "left arrow") for a particular number of update cycles to the corresponding unit in the input image. These external inputs supply a constant activation of 0.25 for the time the external stimulus is present.

For example, for an SOA of 250 msec the prime node would be activated for 25 update cycles, before the other nodes for the four colored circles would be activated in addition. The number of cycles required from the time the stimulus is applied to the time when the difference in activation of the two decision nodes exceeds the threshold $\theta$ is then taken to be the model's response time for an experimental condition (e.g., if 52 update cycles are required after the SOA, the model's response time is considered to be 520 msec.).

To measure the extent to which a given set of excitatory weight values (for the 9 weight groups G5-G13) fits the human data, we define the goodness of a set of (nine) parameters as

$$G(\underline{p}) = \sum_{c\in\mathcal{C}} \parallel RT_h(\underline{p},c) - RT_m(\underline{p},c) \parallel^2$$

where $\mathcal{C}$ is the set of the 32 experimental conditions (words vs. arrow combined with above/below vs. left/right for 8 SOAs), $\underline{p}$ is a vector of parameter values (in our case, $\underline{p} \in [0, 0.1]^9$ for G5 through G13), $RT_h(\underline{p}, c)$ is the average human response time in condition $c$ for parameter values $p$, and $RT_m(p, c)$ is the model's response time (computed from update cycles via $f$). Clearly, the smaller the $G$ value, the better the fit.

## Simulations and Results

Given $G$, it is possible to systematically vary all nine parameters (G5-G13) in order to find the values $\underline{p}$ that result in the smallest $G(\underline{p})$ for the given architecture. These parameter values then define the "best model", i.e., the model that deviates the least from the human data (among all the considered models). The best model can then be used as base model in two ways: (1) as the basis for explanations for the effects seen in the human data, and (2) as the basis for comparisons with other models with different parameters values for connections that are critical to the hypotheses (e.g., G5 or G9/G10).

In the first case, the degree of deviation of the best model from the human data (i.e., the magnitude of $G(\underline{p})$) has to be sufficiently small, i.e., it has to meet some a priori criteria to be considered a good enough fit to be appropriate for generating explanations (e.g., all model data completely within the 95% confidence interval of the human data). In the second case, the goodness of alternative models compared to the goodness of the best model can be used as an indication of the extent to which the particular relationships among parameter values in the best model are critical for the explanation of effects (e.g., the extent to which G5 connections are necessary).

Systematic search of the parameter space yields a minimum of $G(\underline{p}) = 11128$ (for G5=.005, G6=.032, G7=.0325, G8=.04, G9=.035, G10=.029, G11=.065, G12=.039, G13=.0545). This means that the models data points differ by at most 19 msec on average from the human data (this is at most 2 update cycles in the model runs, hence very close to the minimum temporal resolution of the model of 10 msec). The comparison of the model results with the human data depicted in Fig. 4 shows that the model matches the human data – the above/below condition is faster than the left/right condition for words, and the word cue condition is overall slower than the arrow cue condition, which shows no effect of cued location. In particular, the fit for arrows is very close to perfect. Most of the model's deviation comes at two places: (1) in the low SOAs (250 and 500 msec) the model's difference between above/below and left/right in the word cue condition is not as pronounced as in the human case, and (2) the model does not replicate the deviation of the human data for the 1000 msec data point in the above/below condition for word cues, which, upon closer examination, turned out to be an artifact of one outlier subject and is statistically not significantly different from the data point predicted by the model. Moreover, all model data points are within 95% confidence intervals of the human data, so we consider the model appropriate as base model.

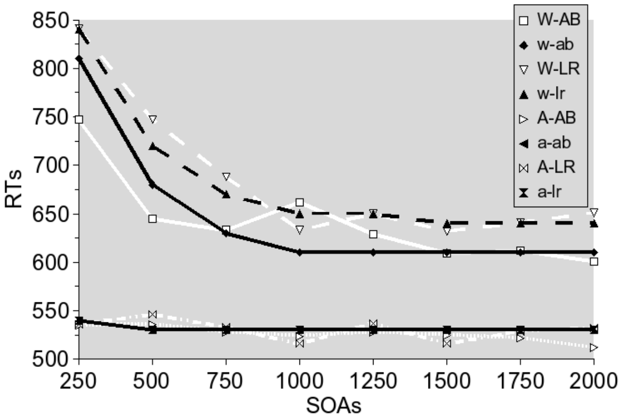

Figure 4: Comparison of the results from model simulations and human data for the 32 conditions (see text for explanation).

The base model directly confirms the first hypothesis (H1) that the differences between above/below and left/right in the word conditions are due to the difference in G9 and G10 connections. Note that in the context of the model, the G9 and G10 connections are the only causes for a difference between above/below and left/right word conditions (even though it might be possible the differences are partly due to factors not captured within the model). This can been from the fact that the goodness of models with identical G9 and G10 is worse than that of the best model (236.68 for models with G10:=G9 of the best model, or 222.68 for models with G9:=G10 of the best model; other models are even worse).

To confirm hypothesis (H2) that differences between all word and all arrow conditions are due both to the direct connections G12 of arrows to the concepts *and* to priming of the visual workspace via G5, we examined the "reduced model" resulting from the best model by setting G5=0. The goodness of that model, 83668, is not only much worse than that of the best model, but the reduced model also effectively eliminates the difference between words and arrows. And while it is possible to obtain a goodness of 41488 for a model G12=0.03 (which is *lower* than the value of the best model for G5=0), this model does not show the human data's characteristics when arrow cues were shown at short SOAs (for space reasons we cannot include additional graphs). Hence, G5 is critical for the effect. This, however, does not make the case yet that G12 is actually needed. Hence, we also consider the model obtained from the best model by setting G12=0, which has a really poor goodness of 915808. Adjusting G5=0.054, we can improve the goodness significantly to 37008 (while keeping G12=0), however, this model also suffers from strong deviations from the human data when arrow cues were shown at short SOAs. Hence, both connections, G5 and G12, are needed, confirming hypothesis (H2).

## Discussion

By confirming (H1), the computational model provides an alternative explanation of the cued location effect that does not involve reference frames. Rather, the difference in RTs is due to a lexical-semantics mapping where the weights between words for directions and direction concepts are different for different directions. Specifically, the weights for "left/right" are lower than those for "above/below". We hypothesize that this difference is due to the learned difference in the validity of word-concept mapping: whenever the words "above" and "below" are encountered, they always denote the directions "above" and "below", whereas the words "left" and "right" sometimes, depending on context, can denote the opposite direction. Thus, from a statistical learning perspective, one would expect the weights between the words "left" and "right" and their corresponding concepts to be lower due to these "inconsistencies" than if the words "left" and "right" *always* denoted the direction concepts "left" and "right".

While this explanation of the cued location effect in terms of statistical learning is consistent with explanations based on spatial reference frames, it is conceptually simpler and based on a general mechanism that is not specific to visual attention tasks. Moreover, the computational model demonstrates that no specific computational mechanism (e.g., one that would effect a mapping between different spatial reference frames) is needed to explain the empirical findings. It should be emphasized, however, that this alternative explanation does not dismiss spatial reference frames as explanatory concepts for other effects (or even this effect, for that matter). Rather, the existence of a simpler explanation might point to need for new experimental paradigms with more sensitive measures to disentangle processing mechanisms that do not involve reference frame from those that do.

By confirming (H2), the model provides an explanation for the cue type effect which suggests that word and arrow cues can bias the selection of visual information in two distinct ways. The first pathway is unique to arrow cues and involves the direct activation of spatial locations within the visual workspace. The existence of this direct activation of spatial locations via arrow cues is theoretically important and may explain recent findings suggesting that arrow cues can elicit reflexive shifts of spatial attention (Gibson & Bryant, 2005). The second pathway is shared by both words and arrows and involves the top-down activation of directional attention.

## Conclusion

In sum, two important conclusions can be drawn from the present study. First, we have provided an alternative, simpler explanation by way of a computational model for the cued location effect observed in the word condition. And second, we have provided new empirical and computational evidence that word cues and arrow cues can bias the spatial selection of visual information in two distinct ways.

## References

Carlson, L. (2003). Using spatial language. *The psychology of learning and motivation*, *43*, 127–161.

Carlson, L., West, R., Taylor, H., & Herndon, R. (2004). Neural correlates of spatial term use. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 1391–1408.

Gibson, B., & Bryant, T. (2005). Variation in cue duration reveals top-down modulation of involuntary orienting to uninformative symbolic cues. *Perception & Psychophysics*, *67*, 749–758.

Gibson, B., & Kingstone, A. (in press). Visual attention and the semantics of space: Beyond central and peripheral cues. *Psychological Science*.

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

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

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

McClelland, J. L., & Rumelhart, D. E. (1988). *Parallel distributed processing* (Vol. 1 and 2). Cambridge: MIT Press.

Posner, M., Snyder, C., & Davidson, B. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*, 160–174.

Scheutz, M., & Eberhard, K. (2004). Effects of morphosyntactic gender features in bilingual language processing. *Cognitive Science*, *28*, 559–588.