Brain & Language 112 (2010) 202–213

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Spatial language, visual attention, and perceptual simulation

Kenny R. Coventry [a,b,*], Dermot Lynott [c], Angelo Cangelosi [d], Lynn Monrouxe [e], Dan Joyce [f], Daniel C. Richardson [g]

[a] *Cognition and Communication Research Centre, Northumbria University, Newcastle upon Tyne, UK*
[b] *Hanse Institute for Advanced Studies, Delmenhorst, Germany*
[c] *School of Psychological Sciences, University of Manchester, Manchester, UK*
[d] *School of Computing, Communication and Electronics, University of Plymouth, Plymouth, UK*
[e] *School of Medicine, Cardiff University, Cardiff, UK*
[f] *Institute of Psychiatry, Kings College London, UK*
[g] *Department of Psychology, University College London, London, UK*

ARTICLE INFO

Article history:
Accepted 1 June 2009
Available online 6 August 2009

Keywords:
Spatial language
Visual attention
Perceptual simulation
Motion processing

ABSTRACT

Spatial language descriptions, such as *The bottle is over the glass*, direct the attention of the hearer to particular aspects of the visual world. This paper asks how they do so, and what brain mechanisms underlie this process. In two experiments employing behavioural and eye tracking methodologies we examined the effects of spatial language on people's judgements and parsing of a visual scene. The results underscore previous claims regarding the importance of object function in spatial language, but also show how spatial language differentially directs attention during examination of a visual scene. We discuss implications for existing models of spatial language, with associated brain mechanisms.

1. Introduction

Spatial language descriptions, such as "*The bottle is over the glass*", constitute an important part of adult language. They also direct the attention of the hearer to particular aspects of the visual world. To do so, they must be grounded in spatial representations of the world; spatial prepositions such as *in*[1] need to be linked to concepts such as INNESS (in some non-linguistic form), which can then be compared to the spatial relations in the world that they describe.

The notion that language drives attention to particular parts of the spatial world is well established. In eye tracking studies, it has been shown that the eyes look towards the objects mentioned as language unfolds (e.g., Allopena, Magnuson, & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) and that such looking behaviour can be anticipatory rather than reactive (e.g., Altmann & Kamide, 2007; Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Knoeferle & Crocker, 2006). In this paper we consider how attention is allocated in a visual scene following a spatial description, and to what extent attentional patterns driven by spatial expressions are regular across specific prepositions and situations.

2. Spatial language and visual attention – two views

There are two views regarding how spatial language directs attention to a visual scene. The first view of spatial language, what we will label SL_{view1}, assumes that the function of spatial language is to narrow the visual search for an object that the hearer is trying to locate (Landau & Jackendoff, 1993; Talmy, 1983). Thus spatial language first directs the attention of the hearer to a reference object (hereafter, RO) in the array being described and then it specifies how attention should be subsequently switched (assuming a serial model of visual attention) from the reference object to the object to be located (hereafter the located object, LO). So "*The X is above the Y*"[2] locates *X* in relation to *Y*, and the direction in which *X* is located with reference to *Y* is specified by the spatial preposition *above* and its associated non-linguistic ABOVE representation.

Regier and Carlson (2001) have proposed an implementation of SL_{view1}, the Attention Vector Sum (AVS) model, that grounds spatial language understanding directly in attention, consistent with evidence of population vector encoding in several neural subsystems (Georgopolous, Schwartz, & Kettner, 1986). The AVS model returns

[*] Corresponding author. Present address: Cognition and Communication Research Centre, School of Psychology and Sport Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK. Fax: +44 191 227 4515.
E-mail address: kenny.coventry@northumbria.ac.uk (K.R. Coventry).

[1] From here on words are italicised, concepts are in upper case, and the objects or relations to which concepts/words refer are underlined.

[2] In order to most clearly expose differences between these views, we will focus on the so-called 'projective' spatial prepositions that have received the most attention in the spatial language literature; *over*, *under*, *above* and *below*.

0093-934X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.bandl.2009.06.001

acceptability judgements for scenes involving simple two-dimensional shapes. An attentional beam is first focussed on the RO at the point that is most closely vertically aligned with the LO, resulting in a distribution of attention across the RO. Additional vectors are defined that are rooted at positions across the RO and that point to the LO. The result is a population of vectors, weighted by the amount of attention at their roots. Summing across this population of weighted vectors produces an orientation that can be compared to the upright vertical (in the case of *above*).

So $\mathrm{SL}_{\mathrm{view1}}$ assumes that most of the early stages of attention are given to the RO in a spatial array, and that attention is then directed to the LO driven by the conceptual relation specified by the preposition. Moreover, the objects in spatial expressions are schematised and do not contribute much to the semantics of a spatial expression in contrast with the fine-grained properties of objects encoded for shape terms, mapping onto the classic neurophysiological distinction between the "what" (ventral) visual pathway and the "where" (dorsal) visual pathway (Landau & Jackendoff, 1993).

A broader view of the function of spatial language, hereafter $\mathrm{SL}_{\mathrm{view2}}$, has been proposed by Coventry and Garrod (2004) in the 'functional geometric framework'. This view, which in some respects is a development of $\mathrm{SL}_{\mathrm{view1}}$, takes into account the way that we experience and use objects in the world. $\mathrm{SL}_{\mathrm{view2}}$ assumes that spatial language comprehension maps onto situation models that provide the most typical relations between objects in the situation in which those objects occur. Consider "*The bottle is over the glass*". According to $\mathrm{SL}_{\mathrm{view1}}$ this sentence would be associated with a minimal representation of a bottle canonically oriented and positioned higher than a glass. In contrast, $\mathrm{SL}_{\mathrm{view2}}$ associates the sentence with knowledge of the typical situation in which the objects are placed where information regarding how the objects typically interact is included.[3] The associated underlying conceptual representation is a perceptual simulation (or 'dynamic kinematic routine'; cf. Coventry & Garrod, 2004) of the typical interaction involving those objects. The symbols in the sentence are mapped onto this underlying representation gleaned from pouring experiences in this case. This view is consistent with the perceptual symbol systems framework (Barsalou, 1999) and with related accounts of language comprehension that involve perceptual "simulations" of events as a key part of language meaning construction (e.g., Kaschak & Glenberg, 2004; Pulvermüller, 1999; Zwaan, 2004). Note that $\mathrm{SL}_{\mathrm{view2}}$ does not discount the importance of geometric processes for spatial description, but rather recognises that a wider range of types of constraints (sometimes treated as pragmatic factors) are at work. These lead to different types of perceptual simulations as a function of the knowledge of situations in which the objects occur. So with "*The bottle is over the glass*", the potential motion of a liquid travelling towards (or otherwise) the glass is important to our understanding of *over* in the prototypical situation in which bottles and glasses occur. In a different context, such as a kitchen tidying context, processing of geometric position may be more relevant than processing pouring for the same expression.[4]

There is now a large body of evidence documenting the importance of a range of so-called 'extra-geometric' variables that affect the comprehension and production of a range of spatial prepositions (see Coventry & Garrod, 2004; Coventry & Garrod, 2005 for reviews). Two key examples will serve us for the rest of the paper.

Coventry, Prat-Sala, and Richards (2001) asked participants to rate how appropriate spatial descriptions containing *over/under/above/below* were to describe pictures of someone holding an object above his/her head. The geometry of the scene was manipulated (the rotation of the held object), and this was crossed with the manipulation of the extent to which the objects (e.g., umbrella, shield, etc.) were shown to be fulfilling their protecting function (see Fig. 1). Ratings for all four prepositions were affected both by the position of the held object and by the success of the object as a protecting object. Rotating the held object away from the gravitational plane reduced the appropriateness of spatial descriptions (consistent with Hayward & Tarr, 1995; Logan & Sadler, 1996). Equally strong effects of the functional manipulation were also present, with the highest ratings for functional scenes (e.g., when an umbrella was shown fulfilling its protecting function by blocking falling rain), and lowest ratings for the non-functional scenes (e.g., when rain was shown to miss the umbrella wetting the person holding it). Moreover, this study found evidence that the comprehension of *over/under* and *above/below* is differentially influenced by geometry and function. Ratings of sentences containing *above* and *below* were better predicted by the geometric manipulation (e.g., position of umbrella in Fig. 1) than ratings for those containing *over* and *under*, while ratings for *over* and *under* were more influenced by function (e.g., position of rain in Fig. 1) than those for *above* and *below*.

In another study, Carlson-Radvansky, Covey, and Lattanzi (1999) asked participants to place one picture of an object *above* another picture of an object. The RO was always an object with a functional part that was dissociated from the centre of the object (e.g., a toothbrush), and the objects to be placed were either functionally related (e.g., a toothpaste tube) or unrelated (e.g., a tube of paint) to the RO. When the RO was presented sideways with the functional part on the left or the right, average placements for the LO were between the functional part and the middle point ('centre of mass', cf. Regier & Carlson, 2001) of the RO. Furthermore deviations towards the functional part were greater for the functionally related object pairs (toothpaste tube and toothbrush) than for the functionally unrelated pairs (paint tube and toothbrush).

These studies, among others, show that "what" objects are and "how" they interact affects spatial language comprehension, contrary to the view that objects are highly schematised for spatial language (Landau & Jackendoff, 1993; Talmy, 1983). However, such effects do not necessarily provide direct evidence against the function of spatial language advocated in $\mathrm{SL}_{\mathrm{view1}}$, and the implications for attention associated with it.

2.1. Testing between $\mathrm{SL}_{\mathit{view1}}$ and $\mathrm{SL}_{\mathit{view2}}$

Given the existence of multiple constraints on the comprehension of spatial language, one needs to establish exactly how these combine to direct visual attention. Carlson, Regier, Lopez, and Corrigan (2006) have shown that the AVS model, consistent with $\mathrm{SL}_{\mathrm{view1}}$, is able to account for at least some of these extra-geometric effects. In the case of a RO with an asymmetrical functional part, such as a toothbrush, it is possible that people simply pay more attention to functional parts of such objects than other parts (see Lin & Murphy, 1997), and the consequences of this for the AVS model are that the vectors rooted in the functional part will be more heavily weighted. Indeed Carlson et al. (2006) show that attention directed from the RO to a LO weighted by object knowledge is able to account elegantly for the differences in placement behaviour reported in Carlson-Radvansky et al. (1999).

[3] Carlson-Radvansky and Tang (2000) have shown that participants rate sentences such as *The mustard bottle is above the hamburger* as more appropriate to describe a scene when the mustard bottle is inverted and is pointing towards the hamburger, rather than when it is canonically oriented. Hence the prototypical spatial relation for these objects in an above situation involves instantiation of a pouring scenario where the LO is required to be inverted to perform its function.

[4] Moreover, for abstract objects without rich situational knowledge to relate them (e.g., *The cross is above the circle*) one might expect geometric simulations to dominate. Much empirical work on spatial language has employed abstract geometric shapes as materials, which may explain the skewed focus on geometric relations in this literature historically. Yet understanding how to interact with the world for the child, and how to describe spatial relations, of course involves objects and object interactions that invariably have meaning for the child. The starting point for spatial language research should therefore involve understanding the mapping between spatial language and more naturalistic materials.

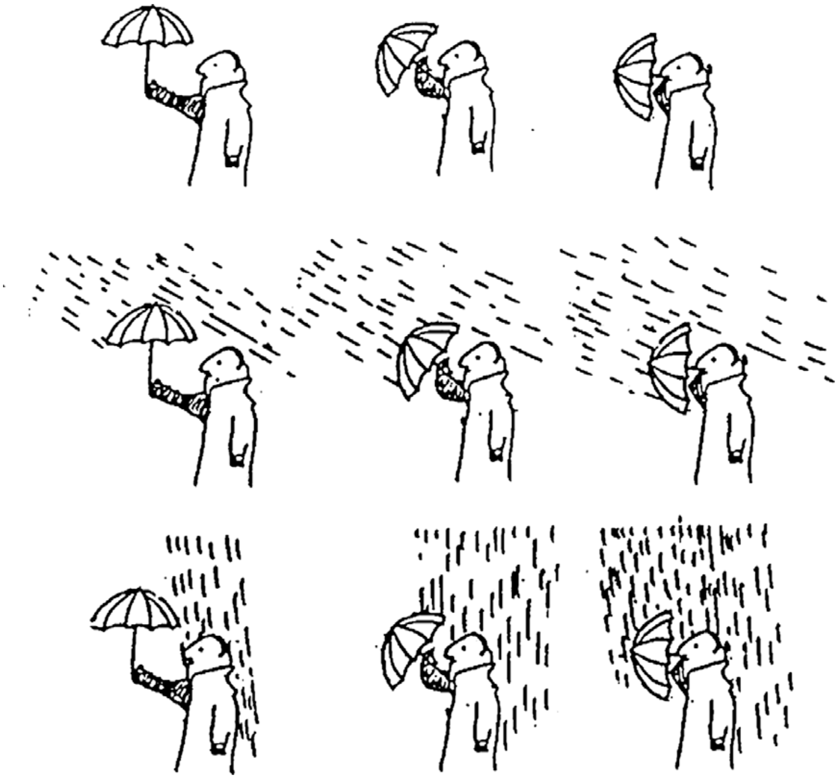

Fig. 1. Examples of scenes used by Coventry et al. (2001). This figure was published in Journal of Memory and Language, Volume 44, Number 3, K.R. Coventry, M. Prat-Sala, L. Richards. The interplay between geometry and function in the comprehension of over, under, above and below, pp. 376–398, Copyright Elsevier (2001).

However, as Carlson et al. (2006) acknowledge, there may be other ways in which function and geometry can be combined. In the materials used by Coventry et al. (2001) functional parts were not misaligned with the centre of the object, and therefore the weighting towards the functional part of the RO in the AVS model does not apply. It is possible, though, that the position of rain in the pictures (e.g., Fig. 1) may cue attention to parts of the umbrella, and that such data can be accommodated within AVS.

An alternative view, proposed by Coventry and Garrod (2004), is that the function of an object is associated with a situational representation of that function, and thus requires a rather different type of perceptual simulation from that of the AVS model. Attention needs to be generated from the rain to the umbrella, or from the toothpaste tube to the bristles on the toothbrush, in order to establish whether the rain or toothpaste will end up in the desired location. This simulation is driven by knowledge of how those objects usually function in familiar situations. Moreover, Coventry and Garrod (2004) speculate that spatial language processing may therefore involve motion processing regions of the visual cortex (middle temporal/medial superior temporal regions) which 'animate' scenes that have implied motion relevant for language judgements. This is in addition to other brain regions, such as the left frontal operculum and left parietal cortices, that have been associated with spatial language processing (Damasio et al., 2001).

The key difference between these two views of spatial language concerns how the position and affordances of objects drive visual attention. In order to discriminate between the accounts of the integration of functional and geometric information, we ran two experiments. The first experiment employed an acceptability rating paradigm using (modified versions of) the materials in Coventry et al. (2001). In order to discount the possibility that the position of the falling objects (e.g., rain) may direct attention to a particular part of the protecting objects (e.g., umbrella) we included two new manipulations. First, rather than showing falling objects either missing the protecting objects or making contact with them (as in Coventry et al., 2001), the falling objects were presented in flight stopping a distance away from the protecting objects, hence never in direct contact with the protecting objects. If the position of the falling objects still affects spatial language judgements, this would provide some preliminary evidence that participants motion process (animate) the falling objects in the scene in order to establish whether the person holding the protecting object will get hit by the falling objects or not, thus affecting language judgements. Second, we compared perfect protecting objects to protecting objects that have exactly the same shape, but are unable to fulfil their functions (e.g., an umbrella with holes in it). This allowed us to keep the position of the falling objects constant, while establishing whether the degree of functional fulfilment is predictive of language ratings. If the presence of holes matters for spatial language ratings, we could be confident that weighted attention to a part of the umbrella cued by the rain (á la AVS) can be eliminated as the only explanation for function effects.

In Experiment 2, we used an eye tracking paradigm to test directly whether participants look at an implied end state of falling liquids/objects in order to return judgements regarding the mapping between language and still images. Eye tracking also begins to unpack exactly how attention is allocated in visual scenes over time, thus affording a useful means of testing the AVS model using real behaviour. For both experiments, it is worth emphasising that the falling objects depicted in all the images are never mentioned in the sentences to be rated. For example, in Experiment 2 the sentences to be evaluated of the form "*The bottle is over the glass*" do not mention any liquid. So effects of the position of the liquid in still images with implied motion would provide strong evidence for $\mathrm{SL}_{\mathrm{view2}}$, consistent with the notion that language comprehension involves recreating the perceptual simulations associated with the situations in which the language was learned.

3. Experiment 1

The first experiment involved three groups of participants. The first group rated the appropriateness of sentences of the form "*The X is preposition the Y*" to describe simple line drawn images. The objects used were always a person and an object with the usual function of protection. Images manipulated both the relative positions of *X* and *Y* in each picture (geometry) and the position of falling objects (not cued in the sentences to be rated) shown in flight a distance away from the protecting object (see Fig. 2 for examples). The protecting objects were also either presented in complete form (i.e., an umbrella without holes) or incomplete form (i.e., an umbrella with holes, thus compromising its function, without any change in object shape).

Two additional groups were asked for non-linguistic judgements about the same images. The second group estimated the percentage of falling objects (e.g., rain) that were likely to make contact with the person, thus providing an independent assessment of the extent to which protecting objects were fulfilling their function for each scene. The third group rated how plausible the images were. It was possible that ratings of the acceptability of a sentence with respect to an image resulted from the acceptability of the images alone rather than the degree of sentence–picture fit – hence the inclusion of this group.

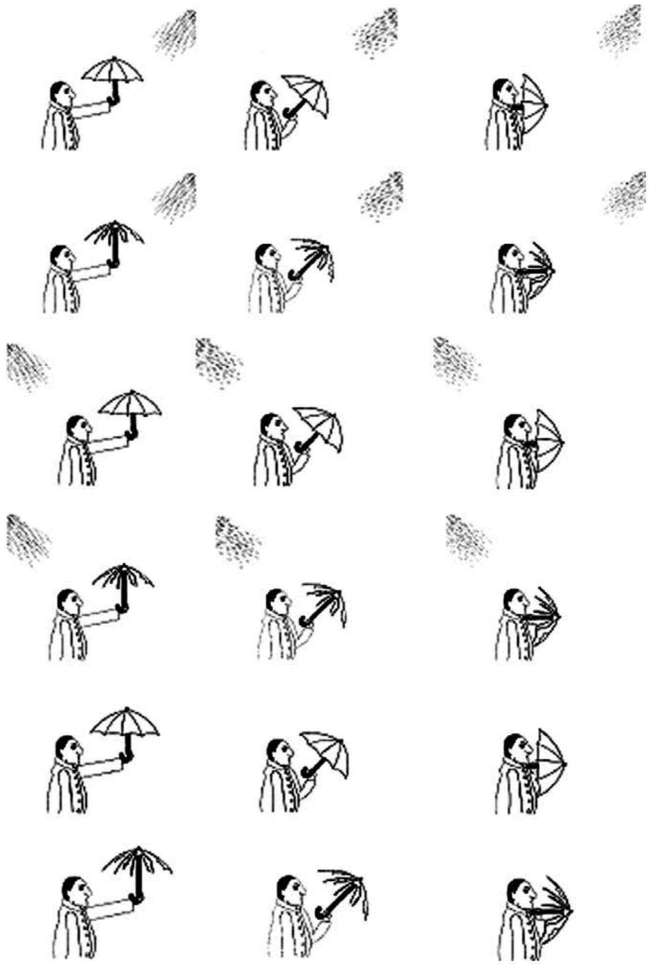

Fig. 2. Examples of materials used in Experiment 1. Panels from left to right show the three levels of position of protecting object. The first two rows show the functional condition, the middle two rows the non-functional condition, and the bottom two rows the control condition (where no falling objects are shown). The odd rows show the complete objects and the even rows show incomplete objects. Sentences presented with these scenes were "*The umbrella is over the man, The man is under the umbrella, The umbrella is above the man, The man is below the umbrella*".

There were several predictions. First, we expected to replicate the results of Coventry et al. (2001) in relation to effects of the positions of both protecting object (geometry) and falling objects (function) on the acceptability of prepositions to describe pictures, plus the differential weighting of these constraints on the appropriateness of *over/under* versus *above/below* (as described above). Furthermore, we predicted a correlation between the percentage of objects expected to make contact with the person holding the protecting object (group 2 judgements) and acceptability ratings for sentences describing the location of the protecting object/person holding protecting object (with a stronger correlation predicted for *over/under* than for *above/below*). Note that such effects would be consistent with the view that participants simulate motion in scenes. We also predicted that object completeness would affect sentence ratings (as well as non-linguistic judgements). Such an effect would strongly suggest that the functional modifications to the AVS model proposed by Carlson et al. (2006) are unable to account for the integration of geometry and function across a range of situations. We did not expect that plausibility judgements would account for the language rating data.

3.1. Method

3.1.1. Participants

Twenty-two participants took part in the sentence acceptability rating study, 15 estimated the percentage of falling objects that would make contact with the person in each scene, and 12 took part in the picture plausibility study. Participants (all English native speakers) were students participating for course credit or payment. None of the participants were aware of the purpose of the experiment.

3.1.2. Design and procedure

Participants in the first group were told that they would be presented with sentences and pictures and that their task was to rate how appropriate each sentence was to describe each picture using a seven point scale (from 1 = totally inappropriate to 7 = totally appropriate), typing their responses using the number keys on the keyboard. The pictures used included two levels of completeness of protecting object (objects with or without holes), three levels of position of protecting object (positioned canonically directly above the other object, at an angle of 45° or at an angle of 90° to the other object) and three levels of position of falling objects/function (the protecting object was expected to either fulfil its function by blocking the falling objects, not fulfil its function, or other objects were not present to make the functional relationship explicit). Given that the end states of falling objects were never shown (the objects in flight stopped a distance before reaching the protecting object) the existence of function effects in this study would only occur if participants project the potential path of the falling objects. The sentences given to rate were presented below each image on a computer screen (e.g., *The umbrella is over the man, The man is under the umbrella*). Four sets of materials were employed; umbrella/man/rain, shield/viking/spears, visor/gardener/spray and hardhat/workman/bricks (making a total of 72 images; see Supplementary materials). The 288 trials (72 pictures × 4 sentences) were fully randomised, and were interleaved with 192 additional filler trials containing different images and sentences testing other spatial relations/spatial language on the horizontal axis (hence *over/under/above/below* did not apply to these scenes). In total there were 480 trials, randomly presented.

The second group of participants estimated the percentage of falling objects that would make contact with the person holding the protecting object. They responded by typing in a number from 0 to 100 where 0% = none of the falling objects would make contact with the person and 100% = all of the falling objects would make contact. These participants were unaware that the study had anything to do with language. Given that the control scenes did not involve any falling objects, these were not given to participants.

The third group of participants rated how plausible each image was, using a seven point scale (from 1 = totally implausible to 7 = totally plausible). This group was instructed to "rate how 'natural' each image is. In other words, rate the extent to which each scene viewed could possibly occur". This group rated all the scenes given to group 1.

3.2. Results and discussion

The mean acceptability rating data (and SDs) for the degree of sentence–picture fit are displayed in Table 1. These data were analysed (collapsed across materials and across superior – HIGHER THAN – and inferior – LOWER THAN – prepositions) using a four-factor repeated measures ANOVA. The variables were preposition set (*over/under* versus *above/below*; motivated by the differential weightings for function and geometry found in Coventry et al., 2001), function (functional, non-functional and control), position of protecting object (canonical, 45° or 90°) and object completeness (holes absent or present). The alpha level was set at 0.05 and follow-up analyses were performed using Tukey (HSD) tests.

Significant main effects of function, $F(2,42) = 23.56$, MSE = 2.58, $p < 0.0001$, position of protecting object, $F(2,42) = 54.04$, MSE = 1.61, $p < 0.0001$, and interactions between function and position of protecting object, $F(4,84) = 5.83$, MSE = 0.262, $p < 0.0001$, position of protecting object and preposition set, $F(2,42) = 36.74$, MSE = 1.49, $p < 0.0001$, and function and preposition set, $F(2,42) = 10.68$, MSE = 0.82, $p < 0.001$ were all present, directly mirroring the results reported by Coventry et al. (2001). Greater effects of function were found for ratings of sentences containing *over/under* than for sentences containing *above/below* while greater effects of the position of the protecting objects were found for *above/below* than for *over/under*.

Of most interest in the present analysis were effects involving object completeness. There was a main effect of object completeness, $F(1,21) = 42.71$, MSE = 2.16, $p < 0.0001$: ratings of spatial expressions for scenes with complete objects were higher ($M = 3.72$) than ratings for spatial expressions for scenes containing objects with holes in them ($M = 3.38$). There was also a significant interaction between function and object completeness, $F(2,42) = 5.42$, MSE = 0.30, $p < 0.01$, and the three-way interaction between preposition set, function and object completeness was also reliable, $F(2,42) = 4.81$, MSE = 0.11, $p < 0.01$: displayed in Fig. 3.

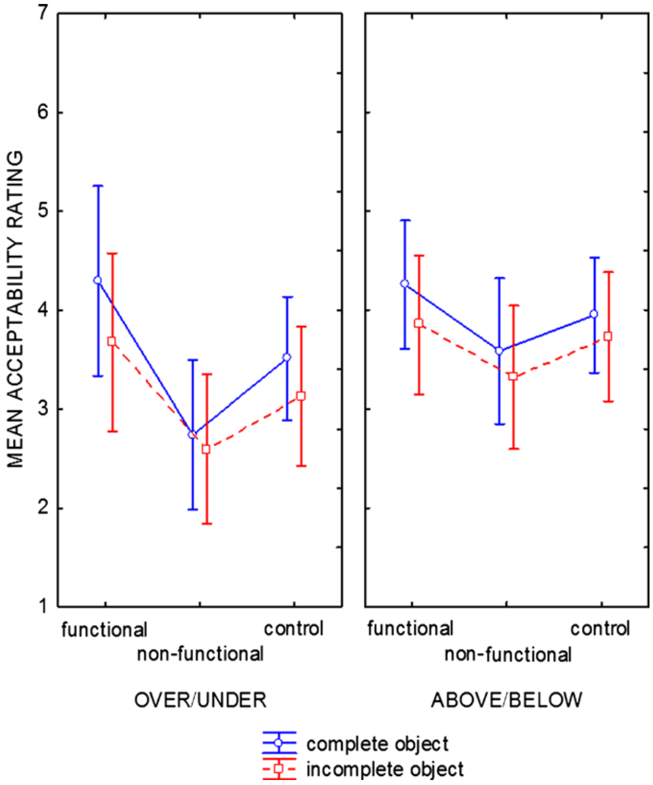

Fig. 3. Interaction between preposition set, object completeness and function in Experiment 1. Error bars indicate 95% confidence intervals.

These interactions reveal that the effect of function on language judgements is mediated by whether the protecting object has holes in it or not. Overall, the effects of function were much more pronounced for the complete objects than for the incomplete objects as expected. However, the three-way interaction shows that the interaction between function and object completeness does not occur uniformly across preposition sets. Running separate analyses for each preposition set, an interaction between function and object completeness was present for *over/under*, $F(2,42) = 6.76$, $p < 0.01$, but absent for *above/below*, $F(2,42) = 2.34$, $p > 0.05$. For *over/under* object completeness reduces the degree of protection the umbrella affords, and consequently leads to a reduction in acceptability for *over/under* for the functional ($p < 0.001$) and control ($p < 0.001$) scenes, but not for the non-functional scenes ($p > 0.05$). The pattern for *above/below* is different. There was a small but reliable ($p < 0.05$) decrease in acceptability ratings for the incomplete objects for all three levels of function, but there was no difference in the size of the function effect comparing the complete and incomplete objects. These results suggest that participants may be animating the falling objects to establish whether

Table 1
Mean ratings (and standard deviations) by condition for Experiment 1.

	Canonical		45°		90°	
	Complete objects	Incomplete objects	Complete objects	Incomplete objects	Complete objects	Incomplete objects
Functional						
Over/under	4.22 (1.14)	3.70 (1.16)	4.51 (1.46)	3.81 (1.29)	4.15 (1.47)	3.52 (1.23)
Above/below	5.10 (0.88)	4.70 (1.13)	4.40 (1.17)	3.95 (1.23)	3.27 (1.13)	2.88 (0.83)
Non-functional						
Over/under	2.65 (0.98)	2.69 (0.96)	2.94 (1.13)	2.75 (1.21)	2.63 (1.08)	2.35 (0.94)
Above/below	4.52 (1.11)	4.19 (1.11)	3.64 (1.22)	3.42 (1.31)	2.60 (0.98)	2.35 (0.91)
Control						
Over/under	3.45 (0.76)	3.10 (0.97)	3.95 (1.10)	3.41 (1.21)	3.13 (0.91)	2.89 (0.88)
Above/below	5.00 (0.89)	4.73 (0.92)	4.32 (0.97)	3.88 (1.24)	2.52 (0.95)	2.59 (0.95)

the protecting object protects or not, but that this is more the case, or is weighted more, for *over/under* than for *above/below*. The fact that an effect of object completeness was present for control scenes where falling objects are not shown may indicate that participants are nevertheless imagining objects falling towards the protecting objects.

A second analysis was run to examine the relationship between the two non-linguistic measures – the percentage of falling objects that would make contact with the person (given by the second group of participants) and the plausibility of the images (rated by the third group) – and the mean acceptability ratings for *over/under* versus *above/below* given by the first group. Cronbach's alpha was calculated for each of the three group measures in order to establish whether there was sufficient reliability for each measure prior to running correlations. The values for the acceptability rating group, % falling objects judgement group and plausibility rating group were 0.986, 0.815, and 0.860, respectively, showing high levels of reliability.

Overall plausibility judgements and % falling objects judgements were significantly correlated, $r_{(191)} = -0.45$, $p < 0.0001$, consistent with the view that the degree of protection afforded by the protecting objects was a factor in the degree of plausibility given to the scenes in the absence of language. Turning to the relationship between the linguistic and non-linguistic measures, the correlations between ratings for *over/under* and % falling objects and plausibility rating were $r_{(95)} = -0.78$, $p < 0.00001$, and $r_{(95)} = 0.12$, $p = 0.24$, respectively. For *above/below* the correlations with % falling objects and plausibility ratings were $r_{(95)} = -0.33$, $p < 0.01$, and $r_{(95)} = 0.25$, $p < 0.05$, respectively. This pattern of correlations shows that, although the two non-linguistic group judgements were correlated, the % falling objects measure correlates much more strongly with language judgements than the plausibility judgements.

Overall the results of Experiment 1 have produced two key findings. First, the degree to which an object protects a person from falling objects (not mentioned in a sentence) affects spatial language comprehension even when the end state of falling objects is not made explicit. We can speculate that participants animate the falling objects to establish whether the protecting object is fulfilling its function or not. Second, the manipulation of the completeness of the protecting object provides evidence that spatial language comprehension is underdetermined by weighting of attention from points on a RO to a LO, as proposed in the AVS model; the shape of the protecting objects remained constant in this experiment, but adding holes still affected judgements. Specifically the addition of holes to an object affects the likelihood that falling objects will make contact with the person holding the protecting object, and hence the degree to which spatial descriptions are rated as appropriate to describe those images.

We can be confident that the language data reflect the fit between language and spatial scenes rather than the plausibility of the pictures alone as the correlations between plausibility and language ratings were only reliable for *above/below*, and weaker than correlations between acceptability ratings and the % of falling objects for all four prepositions.

Some intriguing differences have also emerged between prepositions – *over/under* versus *above/below* – building on earlier findings (Coventry et al., 2001; see also Tyler & Evans, 2003). These could suggest that *above/below* are more dependent on AVS-type calculations in order to affect their judgements, while *over/under* are more affected by motion processing of the falling objects moving towards the protecting object. However, the collection of simple acceptability rating data in this experiment is unable to establish whether (a) specific spatial terms drive attention in visual scenes in different ways, or alternatively whether (b) attentional patterns are identical for all terms, with individual prepositions selectively weighting the outputs of multiple sources of information retrieved from a visual scene. The next experiment uses eye tracking in order to begin to identify if and how spatial language drives attention to different aspects of a visual scene.

4. Experiment 2

This experiment also adopted a rating paradigm, but this time sentences were presented prior to pictures in order to establish how looking behaviour during the pictures is affected by the spatial language judgement to be made. The pictures were colour photographs of a container beginning to pour objects/liquid that would either miss or pour into a second container (see Fig. 4 for examples).

We addressed two issues. First we wanted to establish whether participants would produce acceptability ratings for spatial sentences by animating a static scene. We hypothesised that this mental simulation would be revealed by fixations towards the predicted end path of falling objects. We expected that the functional relationship between objects would influence (i) the likelihood of first fixations to functional and non-functional areas of the bottom object of the scene and (ii) the dwell times (overall and first fixation durations) in the functional and non-functional areas. In other words, in functional scenes, participants will be more likely to fixate and fixate for longer in the functional area of the lower object (i.e., the centre), while in non-functional scenes participants will be more likely to fixate towards the non-functional areas (i.e., where the falling objects would be expected to end up, missing the container). Second, we wanted to use eye tracking techniques to investigate the extent to which individual spatial prepositions weight information gleaned from a visual scene versus drive attention across it. We expected that superior (*over/above*)/inferior (*under/below*) prepositions would direct attention preferentially to the top/bottom objects of the scene, respectively, given their associated semantics (consistent with the AVS model). Furthermore, we wanted to investigate whether participants spend more time looking at the end point of falling objects for *over/under* judgements compared to *above/below* judgements, or alternatively whether these terms simply weight outputs from the same visual processes. Finally, we expected appropriateness ratings to follow the same general patterns as those observed above and previously (e.g., Coventry et al., 2001).

4.1. Method

4.1.1. Participants

Seventeen undergraduate psychology students (all native English speakers) completed the experiment for partial course credit.

4.1.2. Materials

The visual stimuli consisted of eight photographed sets of object pairs taken from Coventry, Christophel, Fehr, Valdés, and Herrmann (in preparation). All scenes showed one object positioned higher than, and displaced to the left/right of a second object. Each object pair appeared in one of four positions (vertically displaced: near or far; horizontally displaced: near or far), and in one of three functional relations between the objects (functional, non-functional and control), comprising 12 scenes (see Fig. 4 and Supplementary materials). For functional scenes, the falling objects from the upper object of the scene (e.g., cornflakes from a box) were shown falling at such a trajectory that they would land in the container below. In non-functional scenes, the falling objects were shown on a trajectory where they would miss the object below. Control scenes contained no falling objects. These 12 scenes were further presented with the top object displaced to the left and to the right of the bottom object for each material set, resulting

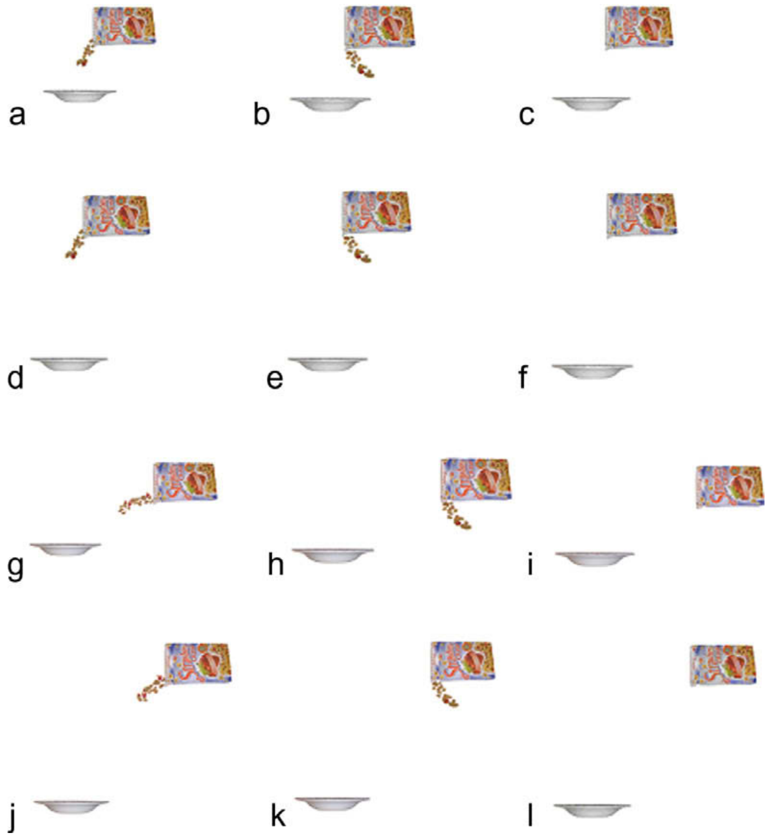

Fig. 4. Examples of scenes used in Experiment 2. Mirror images of all scenes were also used. The rows represent the four positions of the pouring objects on the *x* and *y* axis. The columns represent, from left to right, the functional, non-functional and control scenes. Sentences presented with these scenes were "*The box is over the bowl, The bowl is under the box, The box is above the bowl, The bowl is below the box*".

in 24 scenes for each object pair (in order to alleviate possible attentional biases to the left or right of a scene). Importantly, for each object pair, all aspects of the bottom object remained constant with only the functional relationship between the top and bottom objects being manipulated via the falling objects (i.e., cereal, water, etc. – not mentioned in the sentences to be rated).

For each scene four sentences were constructed of the form "*The X is preposition the Y*", and the preposition slot was filled with *over, under, above* and *below* (e.g., *The jug is over/above the bowl; The bowl is under/below the jug*). The combination of scenes with sentences produced a large number of potential scene–sentence combinations (768 when verbal and visual items are combined; 192 visual scenes × 4 sentences). In order to keep the experiment to an acceptable length for participants, each participant took part in 160 trials (sentence–picture combinations) chosen pseudorandomly from the total scene–sentence combinations; participants saw the full range of materials, but not every material set with every function–geometry–sentence combination. Half the scenes showed the top object displaced on the left; half on the right, and the four prepositions were equally represented throughout the task. The pseudorandom selection was different for every participant, and the order of trails within each selection was fully randomised. Given the length of the experiment filler trials were not added.

For each image, six interest areas were defined (see Fig. 5). Three related to the top object; these were the top object itself (area 1 in Fig. 5) and two areas just below the spout or opening of the object (2 and 3). Objects were shown falling through one of these areas in the functional (2) and non-functional (3) conditions. The bottom object was divided into three horizontal regions. These were the centre area (5) where the falling objects would be expected to land in functional scenes, a near-miss area (6) where the falling objects would be expected to land in non-functional scenes, and a far-miss area (4) the same size as the near-miss area, but on the opposite side of the bottom container where falling objects could not end up. Depending on the analysis being carried out, specific interest areas are incorporated.

4.2. Design

The experiment used a five-factor design; function (functional, non-functional, control), preposition set (*over/under* versus *above/below*), preposition superiority (superior-*over/above*, inferior-*under/below*), distance (near, far) and interest area (described above). The factor of interest area was not used during the analysis of the ratings data. Dependent variables were proportions of first fixations, dwell time for first fixations, total dwell time for each interest area per participant per condition and finally, behavioural data in the form of participant appropriateness ratings for sentence–picture pairs.

Apparatus. An Eyelink II head-mounted eye tracker was used. Participants were unrestrained and seated approximately 75 cm from the 43 cm screen (with screen resolution set to 768 × 1024). The eye tracker recorded pupil position binocularly at 500 Hz which was recorded by the Eyelink host machine. Prior to the experiment and half-way through participants completed a nine-point calibration ensuring that the angle of error was less than 0.6°.

4.3. Procedure

There were two components to this experiment; a sentence–picture rating task that was eye tracked, followed later by a drawing task (that was not eye tracked).

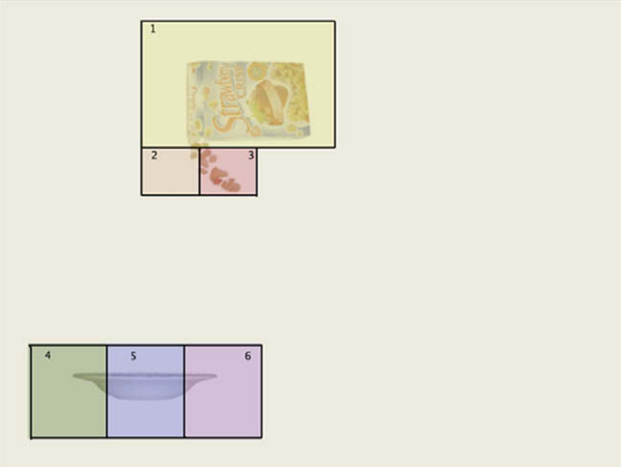

Fig. 5. Interest area regions defined for Experiment 2.

After initial successful calibration of the eye tracker participants were given instructions for the sentence–picture rating task, mirroring those used in Experiment 1. They were told they would be presented with a sentence followed by an image followed by a rating screen. The rating to be provided was to reflect how appropriate the sentence was for describing the image that had followed it. Prior to the experiment proper, participants completed five practice trials. Each trial began with a drift-correction screen; participants pressed the spacebar while fixating on the circle at the centre of the screen. Following drift correction the sentence screen appeared for 2000 ms: sufficient time for participants to fully read the sentence. This was followed by the spatial scene which was displayed for 4000 ms. A rating screen then appeared for 6000 ms (or until a response) with an image of a 1–7 Likert rating scale where participants could input their appropriateness rating by pressing one of the keys on the keyboard number pad (mirroring the scale used in Experiment 1). While eye-movements were recorded for the entire trial, only those during the 4000 ms display of the spatial scene formed part of our eye tracking analyses.

After the eye tracking component, participants completed a short drawing task before debrief. Each participant was presented with a booklet containing 16 scenes which were amongst those presented during the first part of the task. The images contained a mix of functional and non-functional scenes, vertically and horizontally displaced objects, and objects in near and far positions. Participants were simply asked to draw the path of the falling objects (e.g., cereal, water, etc.) from the top object to where they thought they would end up. Based on responses, we established whether participants agreed with our prior classification of scenes as functional or non-functional. In other words, for functional scenes, their drawn trajectories should have the falling objects reaching the bottom object, with trajectories missing the bottom object for non-functional scenes. We found high agreement for vertically displaced scenes (93%) but low agreement for horizontally displaced scenes (23%). Because of this low agreement, all horizontally displaced items were removed from future analyses.

On debrief, it was confirmed that participants were unaware of the true purpose of the study.

4.4. Results and discussion

We first present the results of the behavioural data (acceptability ratings), followed by analyses of dwell times and first fixations to the specific interest areas outlined above.

4.4.1. Acceptability ratings

The mean acceptability rating data (and SDs) for the degree of sentence–picture fit are displayed in Table 2. Rating data were analysed in a 4-factor repeated measures ANOVA. The variables were function, preposition set, preposition superiority and distance. There were main effects of function, $F(2, 15) = 17.81$, MSE = 2.00, $p < 0.001$, and preposition set, $F(1, 16) = 13.04$, MSE = 1.20, $p = 0.002$, consistent with earlier results. There was also a significant interaction between function, preposition set and distance, $F(2, 32) = 4.95$, MSE = 0.39, $p = 0.013$, again mirroring earlier results. Effects of function were more dramatic for sentences containing *over/under* than for sentences containing *above/below*, and conversely effects of distance were more pronounced for *above/below* sentences than for *over/under* sentences.

4.4.2. Dwell times

Dwell time data were analysed in a five-factor repeated measures ANOVA.[5] The variables were function, preposition set, preposition superiority, distance and interest area. The interest area factor had three levels pertaining to the bottom object of the scene; centre, near miss and far miss (regions marked 5, 6, and 4 in Fig. 5). While main effects are somewhat informative, the primary analyses of interest are interactions between interest area and the other factors. More specifically, we focus on whether there are linguistic or functional effects on participants' gaze patterns in the near-miss area of the scenes. If participants are animating scenes, we expected to find longer dwell times for non-functional scenes in the near miss region (area 6) than for functional scenes, and vice versa for the centre region (area 5).

There was a main effect of superiority, $F(1, 16) = 6.57$, MSE = 1762234, $p = 0.021$, and the interaction between superiority and interest area was also reliable, $F(2, 32) = 5.70$, MSE = 1598732, $p = 0.008$. Overall there were increased dwell times for inferior prepositions (*under/below*) compared to superior prepositions (*over/above*), but this was the case only for the central area ($p < 0.05$). There was also a main effect of distance, $F(1, 16) = 15.26$, MSE = 1133,118, $p = 0.001$, and an interaction between distance and interest area, $F(2, 32) = 9.13$, MSE = 708,965, $p = 0.001$; overall there were longer dwell times for scenes with objects near to each other compared to far from each other, which was the case in the central area of the bottom object ($p < 0.001$) and the near-miss area ($p < 0.005$), but not the far-miss area. So when objects are near each other more time is spent looking at relevant regions compared to when objects are further apart, where the eyes have to spend more time travelling between objects.

Of most interest were significant effects and interactions involving interest area, function, and preposition set. There was a main effect of function, $F(2, 32) = 4.54$, MSE = 427,242, $p = 0.018$, with lower dwell times for functional scenes compared to non-functional scenes ($p = 0.059$) and control scenes ($p = 0.017$). There was also a main effect of interest area, $F(2, 32) = 107.74$, MSE = 4,048,435, $p < 0.001$ showing longer dwell times for the centre of the bottom object compared to the near-miss ($p < 0.001$) and far-miss areas ($p < 0.001$). Near-miss areas also received longer dwell times relative to the far-miss areas ($p < 0.01$). So more time is spent looking at regions relevant both in terms of AVS computations – allocated attention from the RO to the LO – and in terms of where falling objects would be expected to end up.

[5] There is some debate regarding the extent to which eye tracking data can be analysed appropriately using ANOVA, given that such data often violate normality (see Jaeger, 2008 for discussion). Examining the distributions, we found that while the fixation data were normal, the dwell time distribution was slightly non-normal. Thus, we log transformed and re-analysed these data finding that all effects reported here were preserved in the new analysis.

Table 2
Mean ratings (and standard deviations) by condition for Experiment 2.

	Near				Far			
	Over	Under	Above	Below	Over	Under	Above	Below
Functional	6.19 (1.11)	6.24 (1.19)	6.21 (1.19)	6.44 (0.87)	5.77 (1.37)	6.11 (1.21)	6.51 (0.69)	6.14 (1.20)
Non-functional	4.48 (1.49)	4.93 (1.61)	5.40 (1.36)	5.61 (1.28)	4.80 (1.60)	5.54 (1.22)	5.60 (1.38)	5.52 (1.53)
Control	5.50 (1.03)	6.00 (0.94)	5.76 (0.90)	5.89 (1.24)	5.84 (1.18)	5.79 (1.39)	6.05 (1.07)	5.96 (1.24)

Notably there was a significant interaction between function and interest area, $F(4, 64) = 3.74$, MSE = 466,705, $p = 0.009$, displayed in Fig. 6. For the near-miss area, participants had longer dwell times for non-functional scenes (205 ms) compared to functional (63 ms) scenes ($p = 0.013$). For the centre area, participants looked longer in the control condition ($M = 2067$ ms) compared to both the functional ($M = 1712$ ms; $t(16) = 2.56$, $p = 0.021$) and non-functional conditions ($M = 1809$ ms; $t(16) = 2.55$, $p = 0.021$).[6] No differences in dwell times were observed for the far miss region.

Finally there was a significant three-way interaction between function, preposition set and interest area, $F(4, 64) = 3.39$, MSE = 569,337, $p = 0.014$, displayed in Fig. 7. The results for the near miss and far miss regions were the same for *over/under* and *above/below*. However, the effect of longer dwell times for the control condition compared to the other conditions for the centre area occurred for *above/below*, but was not significant for *over/under*. These data suggest that *over/under/above/below* are associated with the same attentional allocation vis-à-vis processing of affordances in visual scenes, but that subsequent judgements weight this information differentially across individual terms.

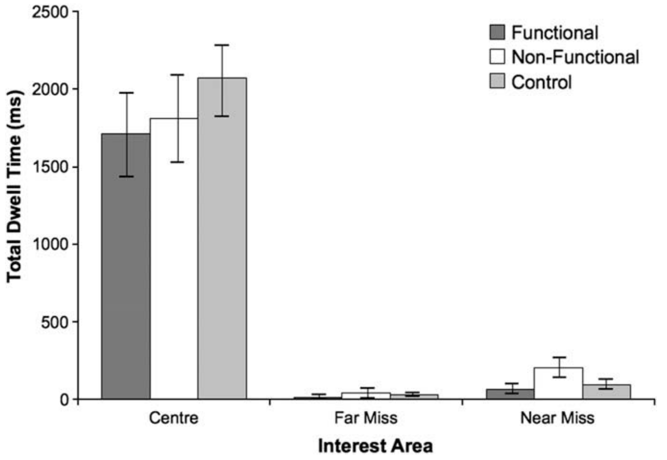

Fig. 6. Interaction between function and interest area. Bars indicate total dwell times per condition. Error bars indicate 95% confidence intervals.

4.4.3. First fixations

Prior to reporting first fixation results based on our predictions for the key interest areas of the bottom object of the scene, we first provide an overview of first fixation results which highlight some regularities in the effects of language on scene processing more generally. First, we analysed time to and duration of first fixations irrespective of interest area, using a four-factor repeated measures ANOVA (function, preposition set, superiority and distance). There was only a significant main effect of distance for both time to first fixations, $F(1, 16) = 43.55$, MSE = 209,214, $p < 0.001$, and duration of first fixations, $F(1, 16) = 24.14$, MSE = 96,282, $p < 0.001$, with time to first fixations being quicker and durations longer for near scenes compared to far scenes.

Second, we examined the proportions of initial fixations made to the top object (interest areas 1–3) or bottom object (interest areas 4–6) of the scene using a five-factor ANOVA (function, preposition set, superiority, distance and interest area).

Overall, there was a main effect of interest area, $F(5, 80) = 439.55$, MSE = 0.31, $p < 0.001$, with all six interest areas being significantly different from the others in terms of attracting initial fixations (all p's < 0.05). The top object in the scene (combining areas 1–3) attracted significantly more first fixations (44%) than the bottom object (10%) (combining areas 4–6), suggesting that people do not always process objects in terms of whether they are a RO or LO, but rather may also process scenes in a top-to-bottom fashion. There was also an effect of superiority, with inferior prepositions leading to more first fixations to the bottom object and superior prepositions leading to a greater proportion of first fixations to the top object, $F(1, 16) = 23.35$, MSE = 0.006, $p < 0.001$ (see Table 3 for proportions). There was an effect of function on proportion of first fixations, $F(2, 32) = 4.45$, MSE = 0.002, $p = 0.02$, with further analysis showing that only control scenes lead to greater fixations to the top object than functional scenes ($p = 0.007$), with no other significant pairwise differences. The factors of distance and preposition set did not affect proportions of first fixations to the top and bottom objects (p's > 0.2).

With respect to our key predictions, we further analysed proportions of first fixations comparing interest areas 5 and 6 of the bottom object: the centre and the near-miss areas. Recall that we predicted non-functional scenes should lead to attention being drawn to the near-miss area of the bottom object. We observed a main effect of distance, $F(1, 16) = 14.85$, MSE = 0.033, $p = 0.001$, with a marginal main effect of function, $F(2, 32) = 2.96$, MSE = 0.022, $p = 0.06$. In planned comparisons, it was found that near scenes resulted in significantly more first fixations to the central area than far scenes ($p < 0.05$), while a greater proportion of fixations were made to the near miss region for non-functional scenes (7%) than for functional scenes (3%); $p < 0.05$ (see Fig. 8). There was no difference in proportion of fixations to the near-miss area comparing functional and control scenes.

In summary, we observed both functional and linguistic effects on participants' allocation of attention, consistent with our predictions both regarding the simulation of falling objects in a scene in

[6] In order to rule out the possibility that strategic effects may account for the data (particularly given the lack of filler items), we ran some additional analyses comparing responses made early in the experiment to those made later in the experiment in order to establish whether patterns are consistent throughout the experiment. This was the case. For example, we performed additional analyses on dwell times for the near miss region of the bottom object, taking responses only from the first 25% (Quartile 1) of trials seen by each participant and then responses only from the final 25% of trials (Quartile 4). Although variance was increased, due to a subsample being selected, the pattern of data was very similar for both early and late responses. In both cases non-functional scenes resulted in longer dwell times than functional scenes ($p < 0.013$ for the first quartile responses; $p < 0.087$ for the fourth quartile responses) and almost equivalent dwell times were found for control and functional scenes for early and late responses (both p's > 0.6). Given that the overall pattern of results holds for responses made early and late in the experiment, we are confident that possible strategising by participants had a minimal effect on these results.

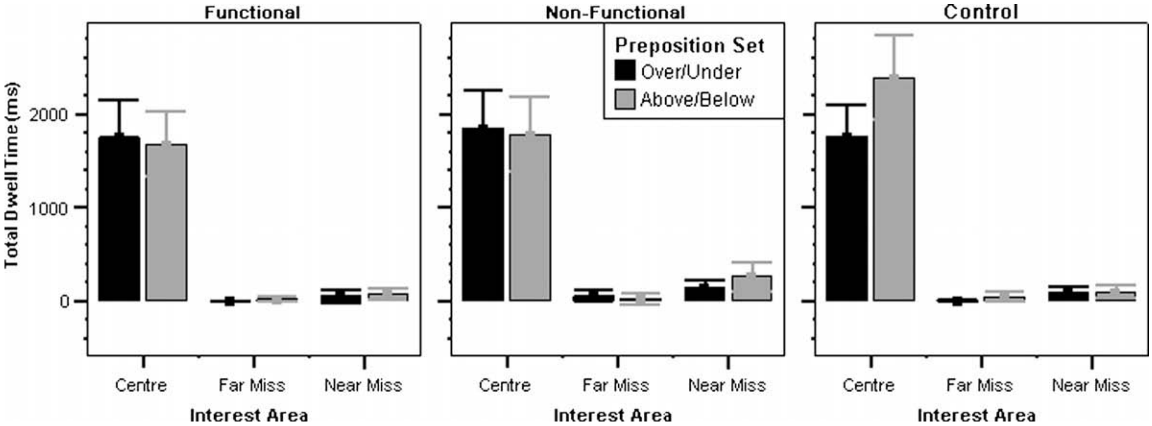

Fig. 7. Interaction between function, preposition set and interest area for dwell times. Error Bars show 95% confidence intervals.

Table 3
Proportion of first fixations (%) to the top or bottom object depending on whether the scene is preceded by a superior or inferior preposition.

	Superior	Inferior
Bottom object	5	14
Top object	45	36

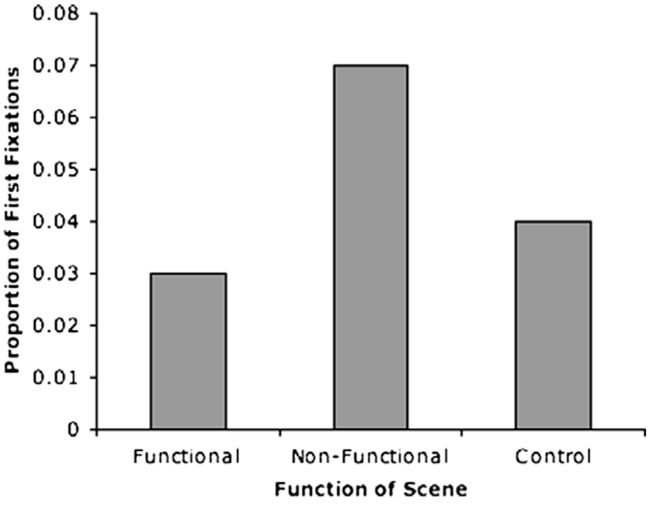

Fig. 8. Proportion of first fixations to the near-miss area of the bottom object (i.e., 1 = 100% of first fixations falling in this area).

line with SL_{view2} and the AVS model regarding how attention is allocated from one object to another object. Where a falling liquid would miss a container if it continued on its path, participants were more likely to look at the area where it would miss, rather than the central region of the relevant object. We also found that inferior prepositions (*under*/*below*) led to increased looking times to the centre of the bottom object relative to superior prepositions, while superior prepositions resulted in increased initial fixations on the topmost object in the scene.

5. General discussion

Across two experiments we considered how spatial language may drive visual attention in visual scenes manipulating both relative positions of the objects named in the sentences in those scenes, and the likelihood with which the expected interaction between objects was afforded. The results of acceptability rating analyses together with the eye tracking data support SL_{view2}, which predicts flexible allocation of attention to a spatial scene as a function of knowledge of how objects interact in situations. In particular, the eye tracking data provide the first concrete evidence that people look at the potential end states of falling objects, and that this is reflected in spatial language judgements. This is consistent with the view that participants are running a simulation of the interaction between the objects where one object is pouring into (or missing) a second container.

The results have implications both for computational models of spatial language comprehension, and for brain regions implicated in spatial language processing. First, modifications of the AVS model proposed to integrate information regarding object function (Carlson et al., 2006) are unable to account for the rating data, or the eye tracking data. This is not to deny the importance of attention allocation from a RO to a LO; the eye tracking data show differences in looking behaviour comparing superior versus inferior spatial terms consistent with the AVS model. Rather, SL_{view2} regards this type of attention allocation as one of a range of visual routines computed from visual scenes as a function of situational knowledge and the specific spatial terms involved.

A different approach to modelling spatial language that shares some of the features of earlier models (e.g., Regier, 1996; Regier & Carlson, 2001), has been developed by Coventry and colleagues (Cangelosi et al., 2005; Coventry et al., 2005; Joyce, Richards, Cangelosi, & Coventry, 2003). The model employs cognitive-functional constraints by extending a series of visual routines (cf. Ullman, 1996) to include operations on dynamic visual input. The input to the model consists of movies, including videos of containers pouring liquids into other containers of the sort we have considered above. A "what + where" code (see Joyce et al., 2003; see also Edelman, 2002) identifies the constituent objects of a visual scene (e.g., RO, LO, and liquid), and consists of an array of some 9 × 12 activations (representing retinotopically organised and isotropic receptive fields) where each activation records some visual stimulus in that area of the visual field. This output is then fed into a predictive, time-delay connectionist network. The network is given one set of activations as input which feed forward to the hidden units. In addition, the previous state of the hidden units is fed to the hidden units simultaneously (to provide a temporal context consistent with Elman's (1990) SRN model). The hidden units feed forward producing an output which is a prediction of the next sequence item. Then, using the actual next sequence item, back propagation is used to modify weights to account for the error. The actual next sequence item is then used as the new input to predict the subsequent item, and so on.

The model provides a mechanism for implementing perceptual symbols (see Joyce, Richards, Cangelosi, & Coventry, 2002), and can "replay" the properties of the visual episode that was learned. So inputting a still image with implied motion, the model is able to establish where falling objects end up. The outputs of this predictive network feed further into a dual-route (vision and language) feed forward neural network to produce a judgement regarding the appropriate spatial terms to describe the visual scene. Thus when seeing a static image with implied motion, the model is able to run a simulation of interaction between objects, based on the mapping between the objects shown in the static scene and past observed interactions with those objects. These simulation results feed forward to language judgements, mirroring the results presented here and elsewhere.

The data and associated model also make predictions regarding brain regions involved in the mapping between spatial language and the visual world. Previously Damasio et al. (2001; see Kemmerer (2006) for a review of related studies with similar findings), using PET scanning, found that both naming spatial relations and naming actions was associated with activity in the left frontal operculum and left parietal cortices, but not in the left infero-temporal cortices (IT) or right parietal cortices (associated with naming objects). In contrast, naming actions was also associated with activity in the lateral temporo-occipital cortices related to motion processing (specifically area MT), but no such activity was found when naming spatial relations. While these results are consistent with SL_{view1} (and Landau & Jackendoff, 1993 in particular), the materials used by Damasio et al. were rather restricted, and they did not systematically vary the prepositions presented with the same scenes.

SL_{view2}, supported by the present results, implicates motion processing regions when mapping spatial expressions to visual scenes when the objects involved occur in situations where motion processing is key for simulation. Perceived motion in humans has been shown to be associated with a cluster of regions in the visual cortex, particularly at the temporo–parieto-occipital junction (MT/MST; Dupont, Orban, De Bruyn, Verbruggen, & Mortelmans, 1994). Kourtzi and Kanwisher (2000) found that there is increased MT/MST activation over baseline when viewing static images which imply movement (e.g., a picture of an athlete in a running pose or a sea wave in mid-air) than when looking at static pictures which do not imply movement (e.g., a picture of an athlete at rest or a calm sea). Although the eye tracking data from Experiment 2 suggest that the extent to which motion processing occurs for such scenes is not different comparing *over/under* versus *above/below*, consistent with SL_{view2} one might expect that different types of language may differentially drive motion processing.

Coventry et al. (in preparation), have examined whether the language judgement to be made when viewing the images presented in Experiment 2 affects the degree to which MT/MST activation occurs. Using a sentence–picture verification task in an fMRI paradigm, they found increased MT/MST activation over baseline when viewing pictures preceded by spatial language (e.g., *The bottle is over the glass*) compared to when the same pictures were preceded by sentences containing comparative adjectives (e.g., *The bottle is bigger than the glass*). Moreover, they also manipulated the degree to which the objects were presented in relative positions affording interaction between the objects. For example, activations for scenes where a bottle was above a glass were compared to scenes with the same objects but with their positions reversed (a glass above a bottle). This manipulation allowed a test of whether a simulation is driven by knowledge of individual objects, or by knowledge of the whole situation in which objects occur. The results were clear; increased MT/MST activation over baseline only occurred when the objects were in the correct positions in which to interact.

Motion processing and associated MT/MST activation should also be present for other spatial prepositions. For example, Coventry and Garrod (2004) have identified 'location control' as the strongest 'dynamic–kinematic' routine, associated with *in* and *on*. An object *x* can be said to be *in/on* an object *y* if object *y* controls the location, or is predicted to control the location, of *x* over time (see also Garrod, Ferrier, & Campbell, 1999; Vandeloise, 1991). This entails that participants should motion process a still image of a pear placed on top of a pile of objects in a bowl when deciding whether "*The pear is in the bowl*" is appropriate to describe that image. This as yet has not been tested.

Coventry and Garrod (2004) also claim that the extent to which different routines apply is affected by context. Intuitively "*The bottle is over the glass*" can mean different things in the context of drinking in a bar versus the context of tidying one's kitchen. Rerunning the fMRI study just described with context presented prior to each sentence and picture could reveal evidence for increased motion processing when viewing a picture of a bottle higher than a glass only when the context makes the dynamic–kinematic routine relevant. We are currently exploring this possibility.

In summary the present results, in tandem with recent neurocomputational modelling work and fMRI studies, provide evidence that spatial language drives attention to visual scenes in different ways as a function of the objects involved and the relations denoted in the sentences preceding those pictures. In line with SL_{view2}, and the 'functional geometric framework', spatial language comprehension is associated with a situational representation of how objects usually function, and thus can invoke a range of types of perceptual simulations, including motion processing where attention is directed to objects not mentioned in the sentence to be evaluated. The full range of potential simulations relevant for spatial language comprehension and the conditions under which they are operable remains to be established.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bandl.2009.06.001.

References

Allopena, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*, 502–518.

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

Cangelosi, A., Coventry, K. R., Rajapakse, R., Joyce, D., Bacon, A., Richards, L., et al. (2005). Grounding language in perception: A connectionist model of spatial terms and vague quantifiers. In A. Cangelosi, G. Bugmann, & R. Borisyuk (Eds.), *Modelling language, cognition and action* (pp. 47–56). Singapore: World Scientific.

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

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

Carlson-Radvansky, L. A., & Tang, Z. (2000). Functional influences on orienting a reference frame. *Memory & Cognition, 28*(5), 812–820.

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.

Coventry, K. R., Cangelosi, A., Rajapakse, R., Bacon, A., Newstead, S. N., Joyce, D., et al. (2005). Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In C. Freksa, B. Knauff, B. Krieg-Bruckner, & B. Nebel (Eds.), *Spatial cognition IV. Reasoning, action and interaction. Lecture notes in computer science* (pp. 98–110). Springer-Verlag.

Coventry, K. R., Christophel, T., Fehr, T., Valdés, B., & Herrmann, M. (in preparation). MT+ activation for static images is driven by knowledge of functional relations and language.

Coventry, K. R., & Garrod, S. C. (2004). *Saying, seeing and acting. The psychological semantics of spatial prepositions*. Hove and New York: Psychology Press Taylor & Francis Group.

Coventry, K. R., & Garrod, S. C. (2005). Spatial prepositions and the functional geometric framework. Towards a classification of extra-geometric influences. In L. A. Carlson & E. van der Zee (Eds.), *Functional features in language and space: Insights from perception, categorisation and development* (pp. 163–173). Oxford, UK: Oxford University Press.

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.

Damasio, H., Grabowski, T. J., Tranel, D., Ponto, L. L. B., Hichwa, R. D., & Damasio, A. D. (2001). Neural correlates of naming actions and of naming spatial relations. *NeuroImage, 13*, 1053–1064.

Dupont, P., Orban, G. A., De Bruyn, B., Verbruggen, A., & Mortelmans, I. (1994). Many areas in the human brain respond to visual motion. *Journal of Neurophysiology, 72*, 1420–1424.

Edelman, S. (2002). Constraining the neural representation of the world. *Trends in Cognitive Science, 6*, 125–131.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Garrod, S., Ferrier, G., & Campbell, S. (1999). *In* and *on*: Investigating the functional geometry of spatial prepositions. *Cognition, 72*, 167–189.

Georgopolous, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 223*, 1416–1419.

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.

Joyce, D., Richards, L., Cangelosi, A., & Coventry, K. R. (2002). Object representation-by-fragments in the visual system: A neurocomputational model. In L. Wang, J. C. Rajapakse, K. Fukushima, S. Y. Lee, & X. Yao (Eds.), *Proceedings of the 9th international conference on neural information processing (ICONP02)*. Singapore: IEEE Press.

Joyce, D. W., Richards, L. V., Cangelosi, A., & Coventry, K. R. (2003). On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. In F. Dretje, D. Dorner, & H. Schaub (Eds.), *The logic of cognitive systems. Proceedings of the fifth international conference on cognitive modelling* (pp. 147–152). Germany: Universitats-Verlag Bamberg.

Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General, 133*, 450–467.

Kemmerer, D. (2006). The semantics of space: Integrating linguistic typology and cognitive neuroscience. *Neuropsychologia, 44*, 1607–1621.

Knoeferle, P., & Crocker, M. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science, 30*, 481–529.

Kourtzi, Z., & Kanwisher, N. (2000). Activation in human MT/MST by static images with implied movement. *Journal of Cognitive Neuroscience, 12*, 48–55.

Landau, B., & Jackendoff, R. (1993). 'What' and 'where' in spatial language and cognition. *Behavioural and Brain Sciences, 16*(2), 217–265.

Lin, E. L., & Murphy, G. L. (1997). Effect of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 1153–1169.

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

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences, 22*, 253–336.

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

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

Talmy, L. (1983). How language structures space. In H. Pick & L. Acredolo (Eds.), *Spatial orientation: Theory. research and application* (pp. 225–282). New York: Plenum 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.

Tyler, A., & Evans, V. (2003). *The semantics of english prepositions: Spatial scenes, embodied experience and cognition*. Cambridge: Cambridge University Press.

Ullman, S. (1996). *High-level vision: Object recognition and visual cognition*. Cambridge, MA: MIT Press.

Vandeloise, C. (1991). *Spatial prepositions. A case study from French*. University of Chicago Press.

Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. In B. H. Ross (Ed.). *The psychology of learning and motivation* (Vol. 43, pp. 35–62). New York: Academic Press.