Event processing in the visual world:

Projected motion paths during spoken sentence comprehension

Yuki Kamide [1], Shane Lindsay [1],

Christoph Scheepers [2], & Anuenue Kukona [1]

[1] School of Psychology, University of Dundee, Scotland, UK

[2] Institute of Neuroscience and Psychology, University of Glasgow, Scotland, UK

Correspondence should be addressed to:

Yuki Kamide

School of Psychology

University of Dundee

Dundee DD1 4HN

Scotland

UK

e-mail: y.kamide@dundee.ac.uk

telephone: +44 (0) 1382 284614

fax: +44 (0) 1382 229993

Abstract

Motion events in language describe the movement of an entity to another location along a path. In two eye-tracking experiments we found that comprehension of motion events involves the online construction of a spatial mental model that integrates language with the visual world. In the first experiment, participants listened to sentences describing the movement of an agent to a goal location with verbs suggesting a more upwards (e.g., "jump") or more downwards oriented path (e.g., "crawl") while concurrently viewing a visual scene depicting the agent, the goal, and some 'empty space' in between. We found that in the rare event of fixating the empty space region between agent and goal, visual attention was biased upwards or downwards depending on the kind of verb. In Experiment 2, the sentences were presented concurrently with scenes featuring a central 'obstruction' which would not only impose further constraints on verb-related motion paths, but also increase the likelihood of fixating the area in-between the agent and the goal. The results from this experiment corroborated and refined the previous findings. Specifically, eye-movement effects started immediately after hearing the verb and were in line with data from an additional mouse tracking task which encouraged a more explicit spatial re-enactment of the motion event. In revealing how event comprehension operates in the visual world, these findings suggest a mental simulation process whereby spatial details of motion events are mapped onto the world through visual attention. The strength and detectability of such effects in overt eye-movements is constrained by the visual world and the fact that perceivers rarely fixate regions of empty space.

Keywords: motion event processing, sentence comprehension, spatial processing, verb semantics, eye movements

Introduction

Over the past few decades, a considerable amount of research has focused on language comprehenders' mental representations of linguistically encoded motion events (e.g., Bergen, Lindsay, Matlock, & Narayanan, 2007; Richardson, Spivey, Barsalou, & McRae, 2003; Verges & Duffy, 2009; Zwaan, Madden, Yaxley, & Aveyard, 2004). This research supports the claim that language comprehension involves *mental simulations* of the events and states that are described in language (Barsalou, 1999; Zwaan, 2004). Specifically, simulation theories assume that our understanding of a linguistically described event recruits the same perceptual and motor representations that would be activated during direct participation in (or observation of) that event.

Uniquely, motion events entail continuous trajectories through both *space* and *time*. Thus, an important question is whether mental representations established during motion event comprehension include detailed information about the continuous paths of moving objects. Given claims about the rich level of detail activated during mental simulations of events (Barsalou, 1999; Zwaan, 2004), simulation theories would suggest that paths can form part of a mental simulation (or mental animation; see Coventry et al., 2013). However, previous investigations have only indirectly addressed this issue. The two experiments reported in this paper investigate the processes by which mental representations of paths implied by verbs are dynamically constructed during motion event comprehension. Contrasting with previous studies, the present experiments examine how perceivers deploy their attention across space and time *on-line* as motion event descriptions linguistically unfold.

Earlier research has used secondary tasks to investigate the influence of linguistically encoded motion events on decisions about visuo-spatial properties *after* participants have finished processing the linguistic stimuli (e.g., Bergen et al., 2007; Richardson et al., 2003; Verges & Duffy, 2009, Zwaan et al., 2004). For example, Bergen et al. (2007) found that the identification of shapes in the upper part of a visual display was slowed following sentences that implied upward motion (and vice versa for downward motion). However, this research does not address whether spatial representations (and mental simulations) occur as soon as the key linguistic input – i.e. the verb – is encountered, or if such representations are part of a post-linguistic, off-line process that involves meta-linguistic decision

making. Similarly, this research does not address whether mental simulations encode specific paths through space, or more generalised spatial biases.

The visual world paradigm is an excellent method for examining online processes during language comprehension, and has also previously been used to investigate comprehenders' spatial representations. In this paradigm, spoken language is simultaneously presented with a related visual display, so that eye-movements on the display can be investigated in close temporal relation to the linguistic input. Hence, by relying on indexes of attention, the visual world paradigm offers a comparatively direct method of examining the construction and updating of spatial representations during language processing, without requiring a secondary task. Previous visual world studies have focused on various aspects of spatial representations, including the updating of spatial representations in motion events (e.g., Altmann & Kamide, 2009), spatial prepositions (Coventry, Lynott, Cangelosi, Monrouxe, Joyce, & Richardson, 2010), and the polysemous use of motion verbs to describe spatial configurations (Richardson & Matlock, 2007).

Recently, Lindsay, Scheepers and Kamide (2013) used the visual world paradigm to study the representation of *speed* in motion events (see also Speed & Vigliocco, 2013). Participants heard sentences with verbs implying fast (e.g., *sprint*) or slow (e.g., *stagger*) movement of an agent to a goal location (e.g., "*the man will sprint/stagger along the path to the house*"). Participants spent more time looking at a visually depicted path in the slow-verb condition, suggesting that the agent was being simulated as spending more time on the path. In contrast, people tended to look earlier at a visually depicted goal (e.g., "*the house*") in the fast-verb condition, suggesting that the agent was being simulated as reaching the goal faster. Importantly, these effects emerged rapidly as the sentence unfolded, suggesting a dynamic process of on-line mental simulation during motion event comprehension

The above visual world studies have primarily focused on how often (proportions of fixations), how long (dwell time), and/or how quickly (latency) listeners look to critical objects in a visual display (e.g., a visually-depicted path). In the current study, our goal was to more directly examine comprehenders' mental representations of linguistically encoded motion events by monitoring the *spatial coordinates* of their fixations over time. Our participants heard sentences with

verbs implying spatial differences in the projected path of motion of an agent towards a goal. Whereas in earlier visual world studies of motion events paths were visually depicted and mentioned in the language (Lindsay et al., 2013), here paths were inferred. In Experiment 1, we presented participants with visual displays that included an agent and goal on the opposite side of the scene (see Figure 1 for an example). We used agentive verbs such as "*jump*" which suggest a motion path in an upwards trajectory through the air to the goal, or verbs such as "*crawl*" which implied contact along the ground and a relatively lower trajectory towards the goal.

Our research aims were as follows. Firstly, we wanted to explore the extent to which projections of a motion path of an event would influence eye movements in the absence of a visually depicted path. A strong version of mental simulation theory (combined with a strong interpretation of the eye-mind hypothesis; Just & Carpenter, 1980) might predict that eye movements actually *trace out* the trajectory of a motion event within the visual field. In this case, participants would fixate along the path of motion through empty space. A weaker claim based on simulation theory is that rather than directly *tracing out* the trajectory of the motion event in empty space, the perceiver's attention (as revealed in overt eye movements) would be biased upwards or downwards in accordance with the described event. If such spatial biases were found, we can further distinguish between two hypotheses: first, that linguistically encoded motion events orient comprehenders' attention upward or downward *globally* (see, e.g., Verges & Duffy, 2009) due to general spatial biases triggered by the verb; or second, that motion events orient comprehenders' attention upward or downward *specifically in relation to the path* of the event. The first hypothesis (global orientation) predicts that verb-related spatial biases in eye-movements should be observable across the entire visual display. The second hypothesis (path-specific orientation) predicts more localised visual attention biases in the spatial region between the agent and the goal. A final aim was to use the temporal precision of eye-tracking to examine the time course of such biases in visual attention in relation to when the critical verb became available in the spoken sentence.

Experiment 1

Method

Participants

Thirty-four native English speakers from the University of Dundee participated for course credit or £4.00. Participants had uncorrected vision or wore soft contact lenses or spectacles, and had no known auditory/visual impairments or language disorders.

Stimuli

Forty-eight visual displays (1024 x 768 pixels) were created using clip-art, and were paired with matching sentences (see Figure 1). Each display consisted of an agent and a goal object. Half of the scenes featured the agent on the left and the goal on the right, and half reversed this positioning. "Alien" agents (introduced to participants as "*Swaliens*") were used to reduce plausibility constraints or directional biases with particular agents, verbs and visual world contexts. Agent colours, shapes and "tentacles" were varied to make 48 unique combinations. Each participant saw each unique agent and goal object once. Scenes were composed of four backgrounds (urban, indoor, outdoor, beach, determined by fixed background colours), with objects chosen to be plausible in each scene. Each scene used a separation of two colours to create a horizontal division behind the entities to represent a horizon (for outdoor scenes) or bottom of a background wall. The latter created a spatially more detailed visual context.

Two types of verbs were used with each scene, so as to suggest a movement of the agent to the goal in an upwards path (Upper Path verbs: *bounce, fly, jump, leap, leapfrog, somersault, spring, vault, bound, float, levitate,* and *pounce*), or a relatively lower path (Lower Path verbs: *crawl, creep, jog, limp, march, roll, stagger, walk, saunter, slide, trek,* and *wander*). Each of the two sentence-versions per scene used the same prepositional phrase (either "*into the X*" or "*onto the X*") to specify the goal. Forty-eight unique two- or three-syllabic names were created for the agents (e.g., "*Foodtock*", "*Neonbert*", "*Paliford*"). Sentences were recorded in mono at 44.1 KHz by a native speaker of English, with the verb phrase cross spliced across sentences.

Each participant saw a further 16 filler items which used verbs of communication, perception or mental state (*complain, contemplate, dream, fantasise, look, shout, stare, think, hallucinate, infer, learn,* and *whisper*, with the propositional phrases "*about the X*" or "*to the X*") with their scenes, so

that participants were not expecting a movement of the agent to the goal in every trial. We further constructed a set of 48 filler items which had an obstructing middle object placed in between the agent and the goal object (with additional agents and goals), with otherwise similar visual properties to the scenes with no obstructions, in order to add variation of the visual configurations of the scenes Obstructions were designed to match the scene, and the directionality of the central obstruction object was balanced, with half facing towards the agent, and half facing away. These items used some of the verbs from the Upper Path and Lower Path conditions, and from the verbs implying no movement (8 sentences for each type). Participants were presented with 72 items in total. Each participant was presented with 32 experimental displays (16 per condition), 16 filler items with verbs implying no movement of the agent, and 24 filler items with a central obstruction. Across participants, all 48 experimental items occurred equally often in each condition.

Procedure

We used an SR Research EyeLink II head-mounted eye-tracker, sampling at 500 Hz from a single eye (viewing was binocular). Participants were told that they would "*see a scene in which there will be a character and some objects. The characters you will see are Swaliens. Swaliens are an alien race and they come in different colours, shapes and sizes. Here is one now, his name is Hillbert*". An example Swalien was depicted. They were further instructed: "*Your task is to listen carefully to each sentence while looking at the screen, and try and understand what will happen in each scene".* On each trial, the scene was presented for a preview period of 1000 ms before the sentence started playing over headphones, and remained on screen until 7000 ms after sentence onset. A drift correct preceded each trial, and the eye tracker was recalibrated every eighth trial. There were four practice trials before the main experimental block. The eye-tracking experiment lasted approximately 20 minutes. The order of items was randomised per participant.

Results

Our analysis of participants' eye movements focused on the y-coordinates of their fixations during critical portions of the sentence. We used an analysis window that spanned verb offset, when

participants had full bottom-up phonological information about the verb, to sentence offset (i.e., during "*_onto_the_sofa*" in "*Foodtock will jump onto the sofa*"; duration *M* = 1366 ms). The 1024 x 768 pixel scenes were divided vertically into three equally sized regions: Agent, Path (middle of screen) and Goal (each region was 341 x 768 pixels in size)[1]. The proportion of looks per region and condition (from sentence onset until 4 seconds later) is plotted in Figure 2. As becomes immediately obvious, the mean percentage of looks to the Path Region, from verb offset to sentence offset, was just about 3% (with only 16 participants contributing data); 52% of fixations were to the Agent Region and 45% to the Goal Region. Due to the limited number of fixations to the "empty" Path Region we increased the analysis window to include fixations up to 500 ms after the end of the sentence, which led to 24 participants and 29 items contributing data.

For these data, we focused our analysis on the mean y-coordinates of fixations in each of the three regions. In all analyses reported below, we reversed the x-coordinates (by subtracting values from 1024) for scenes in which the agent appeared on the right hand side, so that the 'standard' image for analysis had an agent-goal orientation from left to right (see Figure 1). An x- and y-axis coordinate of zero indicates the bottom left pixel coordinate of the standardized image. Figure 3 shows the mean y-coordinates of participants' fixations over time within each of the three scene regions. Curves were computed by averaging the y-coordinates of all fixations within each region into 50 ms time bins. A fixation was scored as 'within' a time bin if (i) its temporal onset was smaller than the temporal offset of the bin AND (ii) its temporal offset was greater than the temporal onset of the bin (trials without a fixation in a given region/bin did not contribute to the mean at that point). In other words, a fixation was counted as being within a bin when it (at least partially) overlapped the bin.

The mean y-coordinates aggregated over the critical time bins per trial (from verb-offset to sentence-offset) are plotted in Figure 4.. A repeated-measures ANOVA with factors Verb Type and Region found a main effect of Verb Type by participants which did not generalise to items, $F_1(2, 44) = 6.6$, $p < .01$, $F_2(2, 48) = 1.8$, $p = .17$, and a significant Verb Type * Region interaction by-items but not by participants, $F_1(2, 44) = 2.3$, $p = .10$, $F_2(2, 48) = 6.9$, $p < .01$. Paired t-tests revealed that in the Path Region there were significantly higher mean y-coordinates both by-participants and by-items for

Upper Path sentences (M = 344, SD = 42) than for Lower Path sentences (M = 304, SD = 45), $t_1(23)$ = 2.23 p= .027, t_2 (28) = 2.17, p = .038, d = .48. There was no such effect of Verb Path for either the Agent or the Goal region ($t_{max} < .5$, $p > .5$). Finally, there was a significant (by subjects) main effect of Region ($F_1(2, 44) = 6.6$, p < .01, $F_2(2, 48) = 2.33$, p = .105.), due to higher fixation Y-coordinates in the Goal than in the Agent or Path region (see Figure 4).

Discussion

In this experiment, we found that participants' overt visual attention tended not to focus directly at the path of a projected motion event. Unlike what might be predicted by a strong version of the mental simulation account (respectively a strong version of the eye-mind hypothesis), the eyes did not "act out" the event described by the language. Instead, fixations appeared to be driven more by the actual presence of objects (agent and goal) and by verbal reference to those objects in the sentence. Fixations on the empty space between the agent and the goal were generally very rare.

Nevertheless, in the few cases where fixations *did* land in the middle Path region during the critical time period, we found a bias for fixating on higher vertical positions in the Upper Path verb condition (e.g., *jump*) as compared to the Lower Path verb condition (e.g., *crawl*). Importantly, these verb-related visual biases appeared to be *specific* to the Path region, i.e. they were not mirrored (at least not as strongly) in either the Agent or the Goal region. Of course, these findings have to be taken with caution due to the low number of data points in the middle Path region. The latter is not only detrimental to power, but also limits generalizability of the present findings (only participants who actually contributed fixations to the middle Path region could reasonably be analysed).

Given existing knowledge of eye fixation behaviour (e.g., Buswell, 1935; Yarbus, 1967; Itti & Koch, 2000), it is perhaps not overly surprising to learn that participants hardly ever fixated on "empty space" regions when there were objects in the visual display to inspect (and taking into account that these objects were explicitly referred to in the linguistic input). While this would argue against a strong link between eye movements and linguistically triggered mental simulation of motion events, eye fixations may still be influenced by the semantics of the verb consistent with the kind of mental simulation processes envisaged here. After all, we did find that when participants launched a

fixation on the "empty" Path region, its y-coordinate was affected by the semantics of the verb, contrasting with fixations on either the Agent or the Goal region.

A major aim behind Experiment 2 was to overcome the paucity of looks to the middle Path region (by encouraging more fixations towards it) and to obtain more conclusive evidence for mental simulation of motion events, and corresponding verb-related biases in fixation coordinates. Specifically, Experiment 2 employed visual displays that contained an intermediate "obstruction" between the agent and goal of the described motion event, such that – if motion paths were inferred – the agent would have to pass either over or under this obstruction in order to reach the goal. Based on the results of Experiment 1, we expected that participants would tend not to fixate directly on regions of "empty space" above or below the central obstruction where the path of the motion event might occur. Instead, despite not being mentioned in the sentences (which just described motion of the agent to a goal object), we expected that participants would fixate on this obstruction object, but crucially, that those fixations would be biased upwards or downwards in accordance with the path of motion implied by the verbs.

Another modification was the introduction of an additional *mouse-tracking* task in Experiment 2, which participants were always asked to complete *after* the main eye-tracking task. The purpose of this mouse-tracking task was to provide an additional measure of the participants' inferred paths of motion in relation to the linguistically described events, and to examine whether these inferred paths of motion could be predictive of participants' eye-movements in the preceding task.

Experiment 2

Method

Participants

Thirty-three participants from the same pool as in Experiment 1 took part. None had participated in Experiment 1.

Stimuli

We modified the materials from Experiment 1 by having 48 scenes with an obstruction version (see Figure 5 for an example), and 24 scenes with no obstruction. Each participant saw 72 scenes: 48 scenes with an obstruction (paired with 32 Upper and Lower Path sentences, plus 16 no-movement fillers) and 24 items without an obstruction (16 Upper and Lower Path sentences, plus 8 no-movement fillers). Here, the scenes without obstructions served primarily as filler items. The mean top of the obstructions was 506 pixels (*SD* = 54 pixels) and the mean bottom was 184 pixels (*SD* = 25 pixels), therefore the mean height of the obstructions was 322 pixels.

In Experiment 2 we reduced the number of verbs from 12 to 8 per verb type, each of which was repeated six times (cf. four times in Experiment 1); for Upper Path verbs we removed *bound, float, levitate,* and *pounce,* and for the Lower Path verbs we removed *saunter, slide, trek,* and *wander*. Due to the removal of verbs there was a different matching of verbs to scenes compared with Experiment 1. Similarly, for the filler items without a movement event we removed four verbs (*hallucinate, infer, learn,* and *whisper*).

Procedure

The procedure was the same as in Experiment 1, except that participants were asked to complete a mouse tracking task after the eye-tracking main experiment. The mouse-tracking task involved exposure to the same stimuli, but in a different random order. Its main purpose was to confirm our intuitions about the trajectories implied by Upper and Lower path items. On each trial, the scene was presented for a preview period of 1000 ms before the sentence started playing. 2000 ms following the end of the sentence, a tone was played and a mouse cursor appeared in the centre of the screen. Participants were instructed to use their dominant hand with the mouse to move the agent in such a way that it performed the action described in the sentence, or if no movement occurred, to press the space bar on the keyboard with their non-dominant hand. Agents were moved by clicking and dragging the agent whilst holding down the left-click button. There was no time limit to each trial, and the task lasted approximately 10 minutes. The order of items in both tasks was randomised per participant.

Results

Contrary to chronological order, results from the mouse-tracking task are presented first, followed by the eye-tracking task.

Mouse-tracking

Data from three participants were lost due to a recording error, leaving 30 participants. Figure 6 shows the aggregated mouse-tracking data across participants for Upper and Lower Path sentences. The plot shows the mean y-coordinates for mouse data points falling within 50 pixel x-coordinate bands along the x-axis. The mouse trajectories confirmed our intuitions: For Upper Path sentences, the mean path was in an approximately parabolic trajectory *over* the middle obstruction (y-coordinate in the Path Region: M = 395, SD = 76), and Lower Path sentences led to the opposite pattern (M = 212, SD = 64) of curving *under* the obstruction to the goal; $t_1(29) = 12$, $p < .001$, $t_2(47) = 22$, $p < .001$.

Eye-tracking data

In this experiment, where an obstruction object was present in the Path region between agent and goal, participants made far more fixations to the Path region than in Experiment 1. Fixations were distributed across the scene as follows: 36% were in the Path Region, 39% in the Agent Region and 24% in the Goal Region. A plot of proportions of fixations per region and condition over time is shown in Figure 7. In the middle Path region, only 6% of fixations failed to land on a rectangle that encompassed the pixels occupied by the middle obstruction, in other words, participants tended not to fixate on the empty space surrounding the middle obstruction (i.e., above, below or to its side).

Figure 8 shows the mean y-coordinates of participants' fixations within each of the three scene regions across time, while Figure 9 shows the mean y-coordinates averaged across the analysis window. Consistent with Experiment 1, we found a main effect of Verb Type, $F_1(1, 29) = 10.6$, $p < .01$, $F_2(1, 47) = 31.0$, $p < .001$, indicating higher y-coordinates for Upper than for Lower Path sentences. The main effect of Region was also significant ($F_1(2, 58) = 62.1$, $p < .001$, $F_2(2, 94) = 10.6$, $p < .001$) due to higher fixation Y-coordinates in the Path than in the Agent or Goal regions (Figure 9), which differs from the pattern in Experiment 1 (Figure 4).. Most importantly, the Verb Type main

effect was further modulated by a reliable Verb Type × Region interaction, $F_1(2, 58) = 12.6$, $p < .001$, $F_2(2, 94) = 42.0$, $p < .001$. In examining this interaction, paired t-tests showed a very clear effect of Verb Type in the Path Region, with Upper Path verbs leading to higher fixation y-coordinates (in pixels, $M = 363$, $SD = 22$) compared to Lower Path verbs ($M = 336$, $SD = 16$); $t_1(32) = 4.8$, $p < .001$, $t_2(47) = 8.8$, $p < .001$, $d = .83$. As can be seen for the Path Region in Figure 8, y-coordinates increased fairly linearly over time in Upper Path sentences (e.g., *jump*) and decreased over time in Lower Path sentences (e.g., *crawl*), with the conditions starting to diverge around verb offset (where the 95% confidence intervals of the means cease to overlap). In stark contrast, and as suggested by Figure 8 and 9, there were no reliable simple effects of Verb Type on y-coordinates in the Agent or Goal Regions (all *ts* < 2).

Finally, we also examined the relationship between effect sizes in the mouse tracking and the eye-tracking task. For each item, and in each task (mouse tracking and eye-tracking), we calculated a measure of effect size, which was the difference between the mean (across participants) Y-coordinates in the Upper and Lower Path conditions obtained for the Path region within the time period from verb onset to sentence offset. Positive scores on this measure mean greater Y-values in the Upper Path condition and negative scores greater Y-values in the Lower Path condition. The by-item correlation between these effect-size scores was small, but significantly positive ($r(46) = .32$, $p < .05$), suggesting that the greater the difference in the mouse tracking task, the greater the corresponding difference in the eye-tracking task.

Discussion

With the presence of many more fixations in the middle Path region, the effects of verb-driven spatial biases in visual attention were much clearer in Experiment 2. Listeners' visual attention was biased more upwards when the verb implied a higher path of motion than a lower one. Unlike in Experiment 1, these spatial biases were primarily due to fixations directed to an obstructing object in the path of the motion event. Another important advance of Experiment 2 is that results (seen in Figure 8) show a clear indication that spatial integration happened during on-line sentence processing, very soon after hearing the verb. Like in Experiment 1, these effects were primarily localised in the

middle Path region, with no clear evidence for vertical biases in the Agent or Goal regions. This suggests that the effects were not just due to global spatial biases induced by the verb (cf. Verges & Duffy, 2009) but instead reflected perceptual simulation process whereby the movement of the agent towards the goal was projected *along the path.* A further interesting finding is that, on an item-by-item basis, effect magnitudes in the second task (mouse-tracking, with a more explicit instruction to 'act out' the described motion events) were to some extent predictive of those in the initial main experiment (eye-tracking), which lends further support to mental simulation of the described events being a contributing factor to the eye-movement results.

General Discussion

Our results are consistent with detailed spatial information being inferred from the linguistic specification of a motion event, including information about its spatial properties. Such results are predicted by simulation-based accounts of language comprehension (Barsalou, 1999; Zwaan, 2004). In line with previous results regarding dynamically updated attention to paths in the visual world paradigm (Lindsay et al., 2013), we see such results as reflecting a process whereby a mental model of a motion event is mapped out onto a static scene, which involves the updating of representations in accordance with a linguistically driven event model. While the visual world provides information on the initial configuration of the event, listeners have to infer and represent the manner and path of the movement event, which incorporates the event's middle and end states, i.e., the trajectory of the motion. This model is "fleshed out" and co-determined by constraints and affordances of the visual-spatial information present in the scene.

Of course, there are several factors that drive eye movements on a visual scene. When hearing a sentence in the visual world paradigm, eye movements are strongly drawn towards the referents mentioned in the sentence (e.g., Tanenhaus et al., 2005). Eye movements may also not be directly yoked to linguistic information and instead result from scene exploration and information extraction. In the present experiments, the middle Path region was not explicitly referred to in the sentence, and the extent to which participants looked at this region was strongly influenced by the presence of a visual object. Without a visual object to fixate upon, participants tended not to look at the Path region

in Experiment 1. With regards to looks to the middle obstruction in Experiment 2, we cannot distinguish in principle between eye movements that were driven by the aforementioned perceptual simulation processes and fixations that were part of scene (or referent) exploration. However, even in the absence of an obstruction in Experiment 1, we still found that fixations in the Path region (and only in that region) were vertically biased in line with verb-specific motion projections, even though these fixations were just landing on empty space. Clearly, the information extracted from a blank region in a scene is different from that extracted while fixating on an object, but our results are consistent in both cases of such eye movements being influenced by the construction of a spatial model of the motion event implied by the sentence. This interpretation is also consistent with how the mouse tracking results indicated how participants construed the paths of the motion events.

The fact that eye movements were biased towards the upper or lower ends of the middle obstruction object in the Path region in Experiment 2 mirrors work on hand-eye coordination (Johansson, Westling, Bäckström, & Flanagan, 2001) demonstrating that when planning actual motion events (moving a bar by hand), people tend to fixate on the edge of obstructions that potentially are in the path, rather than the path in which an object will travel. In our experiment, participants may have construed the motion event to involve contact (e.g., "*bounce*") or avoidance of contact (e.g., "*fly*") of the agent and the obstructions, and in line with the results of Johansson et al. (2001), used fixations on the object itself as a guide to how the agent may avoid (or make contact with) the edges of the object during the described motion event.

There are a number of pressing remaining questions concerning the relationship between mental simulation and language comprehension. One important question is to what extent these results depend on the visual context provided in the visual world paradigm, and how motion events are being processed when there is no spatial frame of reference, as in reading? Are simulations necessary to understand language, even when motion events are not described, and in contexts where visual information is not prioritised?

At present, we certainly cannot fully determine the mechanisms that influence deployment of gaze in the current experiments. However, our findings do suggest the establishment of spatial path representations during on-line sentence processing, in line with the predictions of at least the *weaker*

version of mental simulation theory (cf. introduction). That is, while perceivers did not explicitly *trace out* the described motion events with their eyes (which would assume a strong and immediate link between mental simulation on the one hand and actual eye-movements on the other), their on-line visual attention was nonetheless biased in accordance with verb-specific spatial information, specifically within scene regions that would relate to mentally simulated motion paths.

In conclusion, we have shown that eye movements can be used to help us understand the spatial representations formed during comprehension of motion events. Listeners' overt spatial attention as they inspected a visual scene was influenced in a way consistent with the construction of a dynamic visuo-spatial mental model of that event. These results further our understanding of situated accounts of language comprehension and highlight the interplay between vision and language in how we understand events.

References

Altmann, G., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*(1), 55-71.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Barsalou, L, W. (1999). Perceptual symbol systems. *Behavioural and Brain Sciences, 22*, 577-660.

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

Buswell, G. T. (1935). *How people look at pictures*. Chicago: University of Chicago Press.

Coventry, K. R., Lynott, D., Cangelosi, A., Monrouxe, L., Joyce, D., & Richardson, D. C. (2010). Spatial language, visual attention, and perceptual simulation. *Brain and Language*, *112*(3), 202-213.

Coventry, K. R., Christophel, T. B., Fehr, T., Valdés-Conroy, B., & Herrmann, M. (2013). Multiple routes to mental animation: Language and functional relations drive motion processing for static images. *Psychological Science*, *24*(8), 1379-1388.

Johansson, R. S., Westling, G., Bäckström, A., & Flanagan, J. R. (2001). Eye-hand coordination in object manipulation. *Journal of Neuroscience*, *21*(17), 6917–6932.

Just, MA, & Carpenter, PA (1980). A theory of reading: from eye fixation to comprehension. *Psychological Review, 87*, 329–354

Lindsay, S., Scheepers, C., & Kamide, Y. (2013). To dash or to dawdle: Verb-associated speed of motion influences eye movements during spoken sentence comprehension. *PLOS ONE*, *8*(6), e67187.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64.

Richardson, D, C., Spivey, M, J., Barsalou, L, W. and McRae, K. (2003) Spatial representations activated during real-time comprehension of verbs. *Cognitive Science*, *27*, 767-780.

Richardson, D., & Matlock, T. (2007). The integration of figurative language and static depictions: An eye movement study of fictive motion. *Cognition, 102*(1), 129-138.

Sato, M., & Bergen, B. K. (2013). The case of the missing pronouns: Does mentally simulated perspective play a functional role in the comprehension of person? *Cognition, 127*(3), 361-374.

Speed, L. J., & Vigliocco, G. (2014). Eye movements reveal the dynamic simulation of speed in language. *Cognitive Science, 38*(2), 367-382.

Talmy, L. (2000). *Toward a cognitive semantics, Vol. 1: Concept structuring systems*. The MIT Press.

Verges, M., & Duffy, S. (2009). Spatial representations elicit dual-coding effects in mental imagery. *Cognitive Science, 33*(6), 1157-1172.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.

Zwaan, R. A. (2003). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of Learning and Motivation, 44*, 35-62.

Zwaan, R, A., Stanfield, R, A., & Yaxley, R, H. (2002). Language comprehenders mentally represent the shape of objects. *Psychological Science, 131*, 68-171.

Zwaan. R. A., Madden. C. J., Yaxley. R. H., & Aveyard. M. E. (2004). Moving words: dynamic representation in language comprehension. *Cognitive Science, 28*, 611-619.

Author note

Shane Lindsay is now at the Department of Psychology, University of Hull. Anuenue Kukona is now at the Division of Psychology, De Montford University.

We are grateful for Karl Smith Byrne for their help with data collection for Experiment 1, and thank R Gordon Brown for his help with stimulus creation and data collection for Experiment 2. We gratefully acknowledge support from ESRC Grants RES-062-23-2842 to YK & CS and RES-062-23-2749 for YK.

Figures

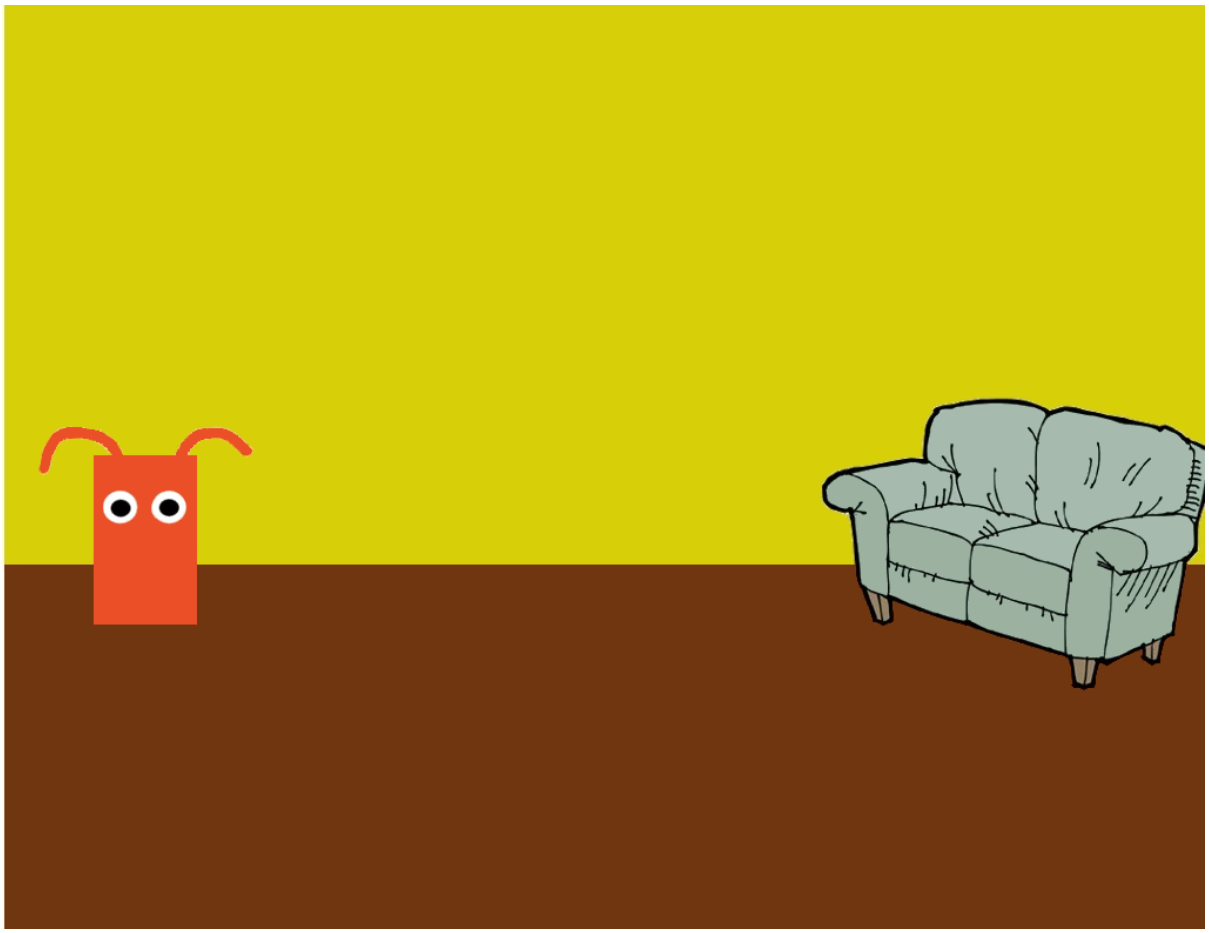

Figure 1. Example visual stimulus for the sentences: “Foodtock will jump onto the sofa” (Upper Path), “Foodtock will crawl onto the sofa” (Lower Path) in Experiment 1.

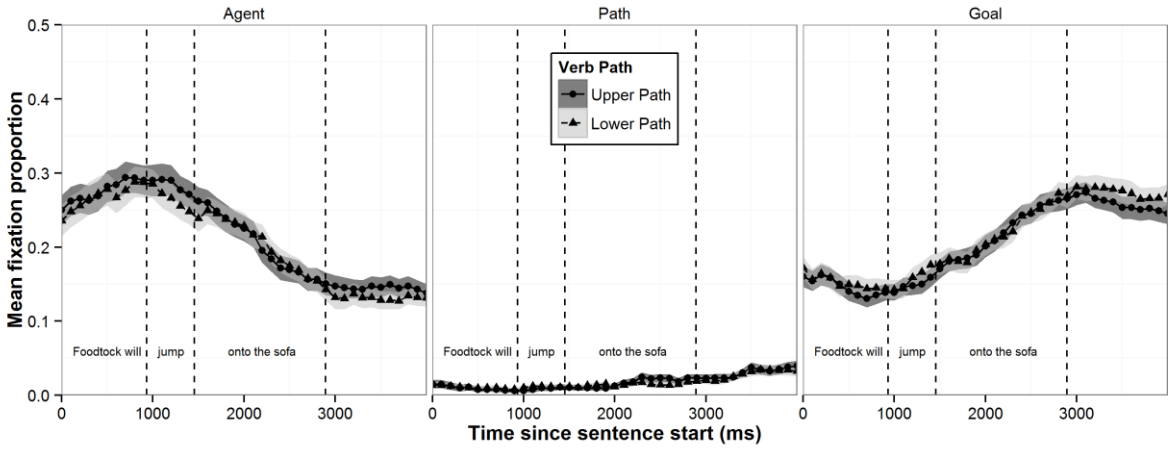

Figure 2. Mean proportion of fixations per scene region and verb condition over time (Experiment 1). broken down by region (Agent, Path, Goal) and verb condition (Upper Path vs. Lower Path). Error ribbons show standard errors.

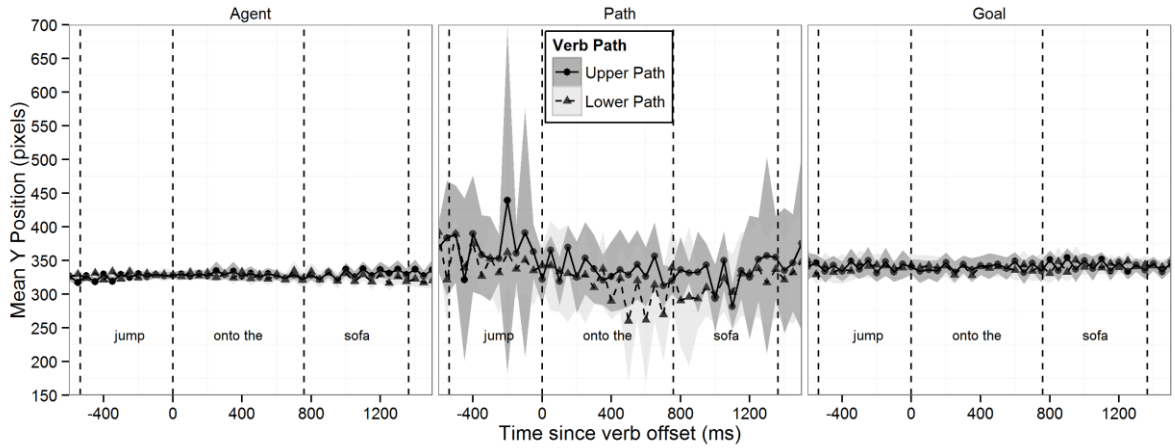

Figure 3. Mean Y-coordinate of fixations over time (zeroed at mean verb offset) in Experiment 1, broken down by region (Agent, Path, Goal) and verb condition (Upper Path vs. Lower Path). Error ribbons show corrected within-participants 95% confidence intervals (cf. Morey, 2008).

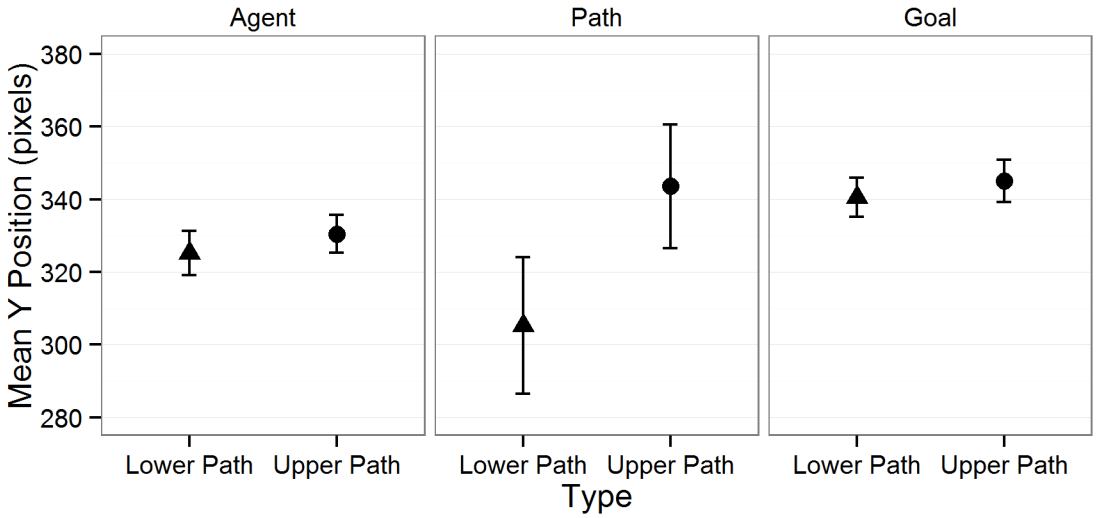

Figure 4. Mean Y-coordinate of fixations (collapsed across the entire analysis window) for Experiment 1, broken down by region (Agent, Path, Goal) and verb condition (Upper Path vs. Lower Path). Error bars show corrected within-participants 95% confidence intervals (cf. Morey, 2008).

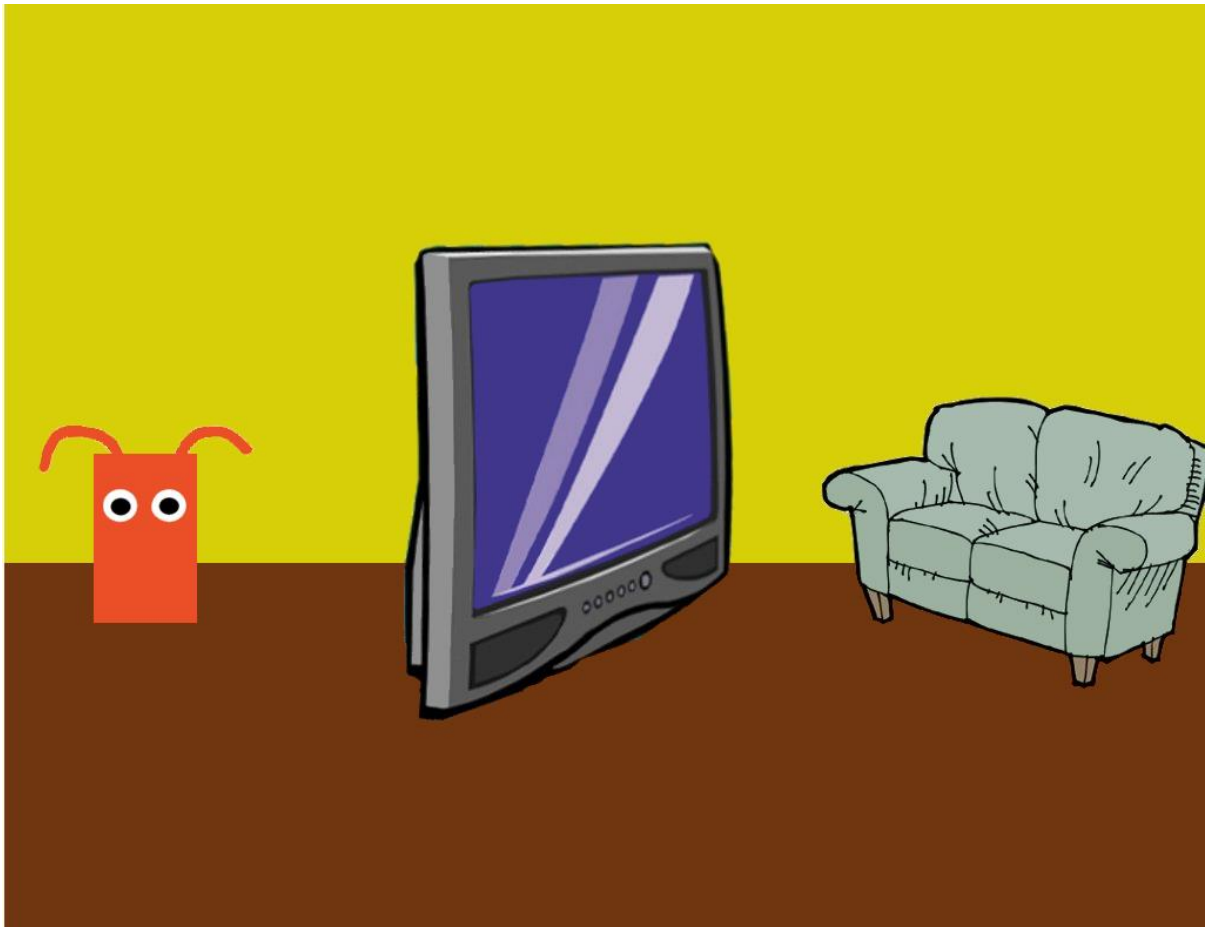

Figure 5. Example visual stimulus for the sentences: “Foodtock will jump onto the sofa” (Upper Path), “Foodtock will crawl onto the sofa” (Lower Path) in Experiment 2.

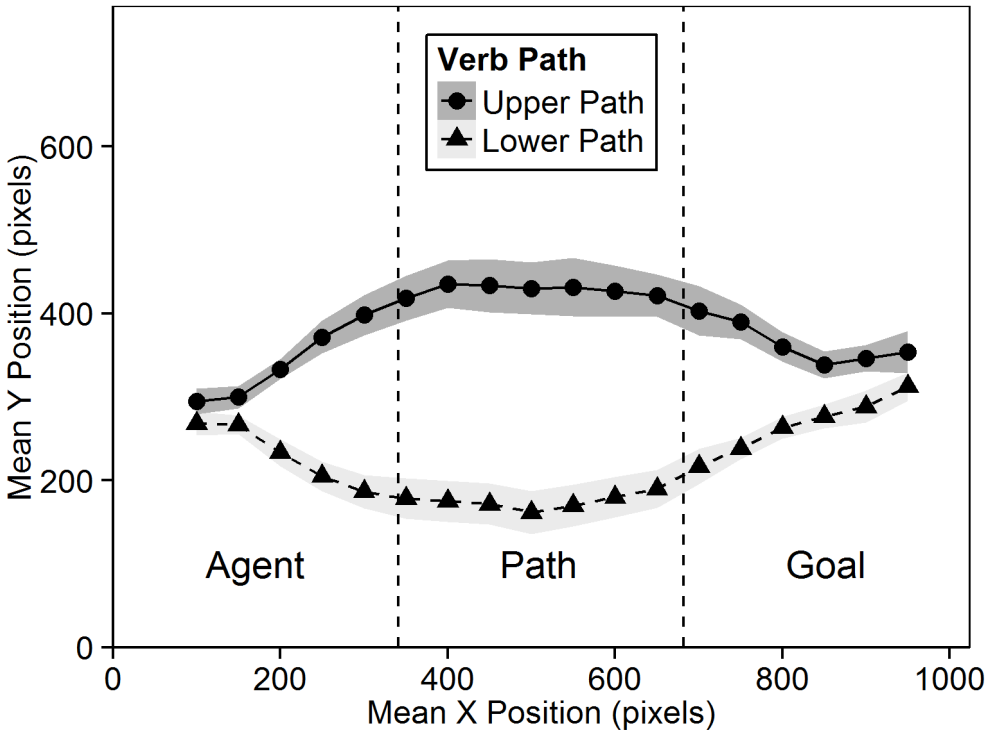

Figure 6. Mean X-Y coordinates of the mouse paths for the Upper and Lower Verb conditions in the mouse tracking task of Experiment 2. Error ribbon shows corrected within-participants 95% confidence intervals (cf. Morey, 2008).

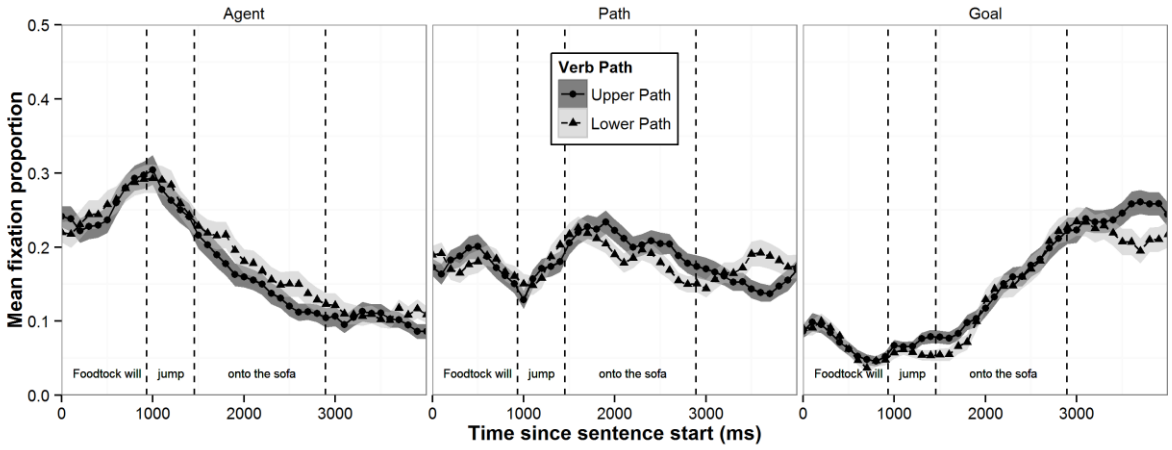

Figure 7. Mean proportion of fixations per scene region and verb condition over time (Experiment 2). broken down by region (Agent, Path, Goal) and verb condition (Upper Path vs. Lower Path). Error ribbons show standard errors.

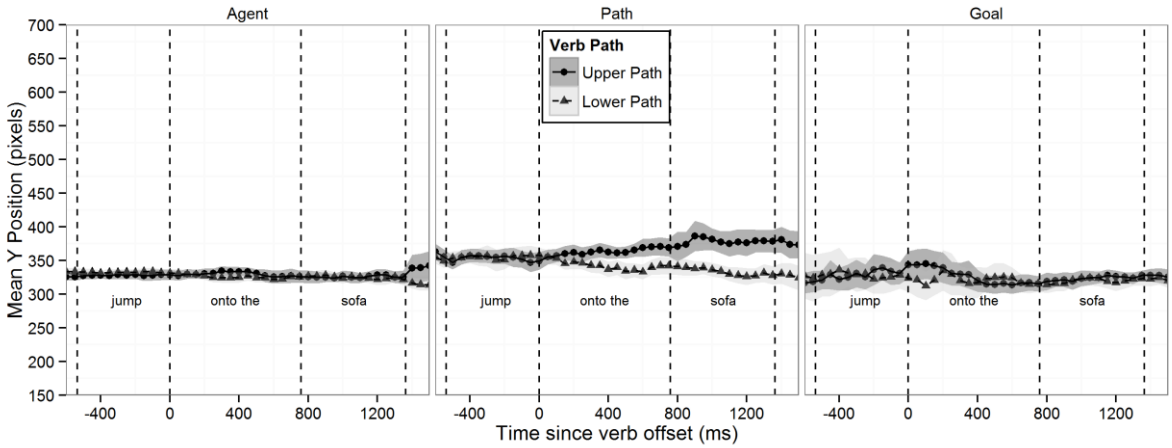

Figure 8. Mean Y-coordinate of fixations over time (zeroed at mean verb offset) in Experiment 2, broken down by region (Agent, Path, Goal) and verb condition (Upper Path vs. Lower Path). Error ribbons show corrected within-participants 95% confidence intervals (cf. Morey, 2008).

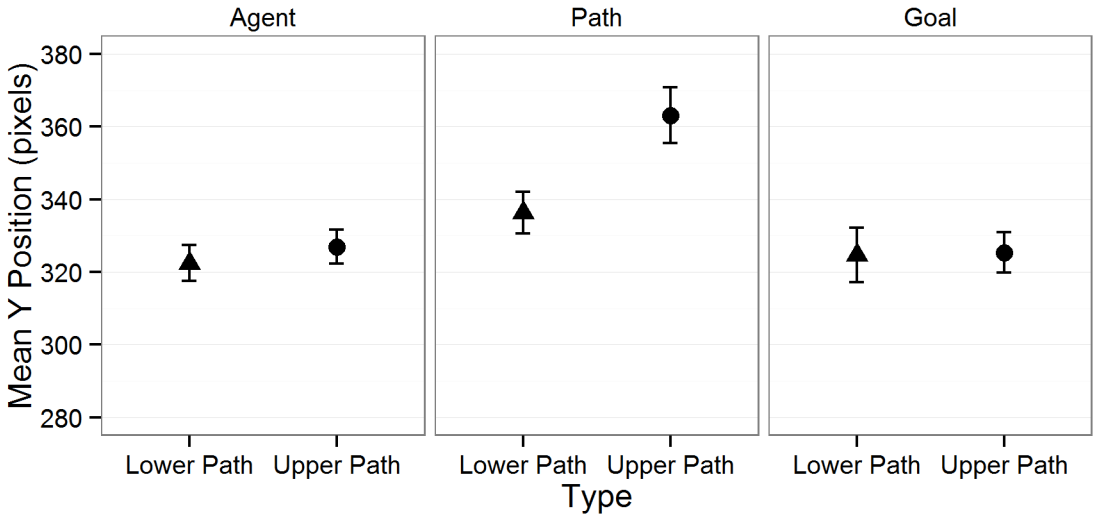

Figure 9. Mean Y-coordinate of fixations (collapsed across the entire analysis window) for Experiment 2, broken down by region (Agent, Path, Goal) and verb condition (Upper Path vs. Lower Path). Error bars show corrected within-participants 95% confidence intervals (cf. Morey, 2008).

Footnote

[1] Within each of these three broad regions, the majority of fixations were within or very close to the pixels occupied by each entity. Using more narrowly defined region of interest formed by a rectangle around the extent of the Agent, Path and Goal objects, and excluding fixations to empty space, produced similar results in eye-tracking to those reported here.