The activation of object-state representations during online language comprehension

Xin Kang[1,2*]

Gitte H. Joergensen[3,4]

Gerry T.M. Altmann[3,4]

[1] Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

[2] Brain and Mind Institute, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

[3] Department of Psychological Sciences, University of Connecticut, Storrs, CT, USA

[4] CT Institute for the Brain and Cognitive Sciences, University of Connecticut, Storrs, CT, USA

Word count: 11309

Corresponding author:

Xin Kang

Email: xin.kang@cuhk.edu.hk

Abstract

Understanding the time-course of event knowledge activation is crucial for theories of language comprehension. We report two experiments using the 'visual world paradigm' (VWP) that investigated the dynamic mapping between object-state representations and real-time language processing. In Experiment 1, participants heard sentences that described events resulting in either a substantial change of state (e.g. *The chef will* ***chop*** *the onio*n) or a minimal change of state (e.g. *The chef will* ***weigh*** *the onion*). Concurrently, they viewed pictures depicting two versions of the target object (e.g., an onion) corresponding to the intact and changed states, and two unrelated distractors. A second sentence referred to the object with either a backward or a forward shift in event time (e.g. ***But first/And then****, he will smell the onion*). In Experiment 2, Degree of Change was manipulated by using different nouns in the first sentence (e.g. *The girl will stomp on* ***the penny/egg***). The second sentence was similar to the ones used in Experiment 1 (e.g., ***But first/And then****, she will look at* ***the penny/egg***). The results from both experiments showed that participants looked more at the 'appropriate' state of the object that matched the language context, but the shift of visual attention emerged only when the object name was heard. Our findings suggest that situationally appropriate object representations do trigger eye movements to the corresponding states of the target object, but inappropriate representations are not necessarily eliminated from consideration until the language forces it.

Word count: 241

Keywords: object-state; mental representation; language comprehension; visual word paradigm

Introduction

World knowledge of objects and events play an important role in language comprehension (Altmann & Ekves, 2019; Altmann & Mirković, 2009; Elman, 2009; McRae & Matsuki, 2009; Kuperberg & Jaeger, 2015). An existing body of work has demonstrated that perceptual properties of objects such as shape (Zwaan et al., 2002), orientation (Stanfield & Zwaan, 2001), motion direction (Zwaan, Madden, Yaxley, & Aveyard, 2004), and visibility (Yaxley & Zwaan, 2007) are accessed more quickly when they match rather than mismatch the language context. These findings are consistent with theories that language comprehension involves establishing mental models of described events and the properties of the objects taking part in those events (Johnson-Laird, 1983; van Dijk & Kintsch, 1983). Furthermore, object properties in these mental models can be updated when changes of location (e.g., from one room to another) or time (e.g., an hour later) are described in the language context (Glenberg, Meyer, & Lindem, 1987; Radvansky, 2009, 2012; Randvansky & Copeland, 2006, 2010; Speer & Zacks, 2005; Swallow, Zacks, & Abrams, 2009; Zacks, Speer, Swallow, Braver, & Reynolds, 2007; Zwaan, 1996; Zwaan & Madden, 2004; Zwaan & Radvansky, 1998).

However, less is known about when and what properties of objects are activated during online comprehension. Consider the following scenario:

The chef will chop the onion. But first, he will smell the onion.

As we read the first sentence, we may draw on our knowledge from the real world to predict that the "*chop*" action separates the existence of the onion into at least two conflicting representational states – being intact prior to the described action and being in several pieces after being chopped. Do we continue to keep both states

activated (i.e. the intact state and the changed state) even when a subsequent sentence cues us to retrieve just one of the states (e.g., *but first*)? If not, at what point as the sentences unfold, do we activate the intended 'interpretation' of the onion and suppress the unintended one? According to Altmann & Ekves (2019), events are "*ensembles of intersecting object histories*" and that the processing of an object that has changed state entails (at least transient) activation of its previous states that may compete for selection.

The current study used the visual world paradigm (VWP), a method that has been used previously to examine the online mapping between event representation and unfolding language (e.g., Altmann & Kamide, 1999, 2007, 2009; Chambers, Tanenhaus, & Magnuson, 2004; Kamide, Altmann, & Haywood, 2003; Kamide, Lindsay, Scheepers, & Kukona, 2016; Knoeferle, Crocker, & Scheepers, & Pickering, 2005; Kukona, Altmann, & Kamide, 2014). Converging evidence suggests that the visual world paradigm (VWP) is sensitive to the dynamic mapping between language and both event and object representations. In this paradigm, eye movements around the visual displays are used as an index of the mapping between mental representations of events and language as it unfolds. Prior research has shown that eye movements are not only drawn to objects that are mentioned or anticipated in the language (e.g., Altmann & Kamide, 1999), but also to entities that are semantically related either in shape (Dahan & Tanenhaus, 2005; Huettig & Altmann, 2007), location (Hoover & Richardson, 2008), category (Huettig & Altmann, 2005; Yee & Sedivy, 2006), function (Kalénine, Mirman, Middleton, & Buxbaum, 2012), or the described situation (Altmann & Kamide, 2007, 2009; Knoeferle & Crocker, 2007; Kukona, Altmann, & Kamide, 2014).

Altmann and Kamide (2009) asked participants to view visual scenes (containing e.g. a table, a bottle of wine, a wine glass (on the floor), a bookshelf and a woman), and listen to sentences such as "*The woman will put the glass onto the table"* (the 'moved' condition) *or "The woman is too lazy to put the glass onto the table"* (the 'unmoved' condition). After listening to one or other of these two sentences, participants heard "*Then, she will pick up the bottle, and pour the wine carefully into the glass*". At "carefully", participants looked at the table more in the 'move' condition than in the 'unmoved' condition. This result demonstrates that object representations can be updated according to linguistically encoded but not visually perceived location changes.

Using functional magnetic resonance imaging (fMRI), Hindy, Altmann, Kalenik, and Thompson-Schill (2012) and Solomon, Hindy, Altmann, & Thompson-Schill (2015) argued that multiple representations of the same object, in its different states as described by the unfolding language, can be simultaneously activated in language comprehension. Participants read either "*The woman will chop the onion. Then, she will smell the onion*" or "*The woman will weigh the onion. Then, she will smell the onion*"; they found increased BOLD response to the former relative to the latter in the posterior left ventrolateral prefrontal cortex (subsuming parts of Broca's area), and specifically in voxels that were sensitive to word-color interference in a Stroop task. They interpreted this effect as indicative of *competition* between the distinct representations of the onion. No such change in BOLD was observed when the second sentence was "*Then, she will smell another onion*" (in which case there was no representational conflict between distinct representations of the same onion). Further, they found that Degree of Change entailed by the description of what

happened to the object, as rated by other participants (e.g. chopping an onion entails greater change than peeling it), correlated with that BOLD response, adding to the evidence that these effects were due to differences in object representation. Due to the limited temporal resolution of fMRI, it remains unclear at what point in time these competition effects occurred – fMRI cannot track across time the activation of the distinct object representations (e.g. the intact state versus the changed state of the same onion). This is where the visual world paradigm (VWP) excels.

The present study explored the construction and modification of event-based object representations as a spoken description of an object changing state unfolded in real-time. In two experiments, participants were asked to listen to two-sentence stimuli while looking at a visual display showing four clipart items (the target object depicted simultaneously in two distinct states and two unrelated distractors). The first sentence established the context of the event when the target object was described as experiencing either a minimal (e.g., "*The chef will weigh the onion*") or substantial change (e.g. "*The chef will chop the onion*"), while the second sentence referred back to this object in the context of a new (minimally changing) action ("*But first/And then, he will smell the onion*"). We use the terms "minimal" and "substantial" following the convention in Hindy et al (2012), given that in many of the experimental items there is some type of change (e.g., *weighing* something causes changes in location).

We examined language-mediated eye movements to measure when people shift their attention between the competing states of the target object, as directed by the unfolding language. We further examined when previously 'outdated' but now 'relevant' information becomes available in online language comprehension, and whether language-mediated eye movements are sufficiently sensitive to allow exploration of

such object-state tracking (contrasting *"And then, he will…*" with "*But first, he will…*" – in the latter case, the initial intact state of the onion becomes the more relevant of the two, whereas in the former case, it is no longer relevant). If perceptual properties of the situationally relevant target state were retrieved as language triggers changes in the representation of the situation, we would expect to see differences in eye movements towards the conflicting states of the target object depending on the language context (that is, we expect to see attention switching from one state to the other – at issue is *when* the switch occurs).

While we can predict with some certainty that for a sentence pair such as "*The chef will chop the onion. But first, he will smell the onion*" participants will look towards the appropriate, intact onion during the final "*the onion*", we cannot be certain *when* these looks will begin to be directed at that onion. In the second sentence, in principle, the earliest we could see sensitivity to the temporal properties of the situation is at, or soon after we hear the part of the sentence that temporally anchors the second sentence with respect to the first sentence. The scene shows two onions, and if participants interpret the depicted chopped onion as the resultant state from "*The chef will chop the onion*", on hearing "*but first*" they would know at that point that the chopped onion is no longer situationally relevant given the temporal reference to an earlier time frame. Thus, if interpretation of the temporal context as indicated by "but first/and then", and of the corresponding object states, is maximally incremental (see Altmann & Mirkovic, 2009, for discussion), and if such interpretation limits attention to (in this case) the visual environment, evidence for such incrementality might manifest as looks away from the situationally-irrelevant (chopped) onion upon hearing "*but first*", reflecting the anticipation that some *other* of the objects in the scene is likely to be

referred to subsequently.

Experiment 1

In Experiment 1, we ask: If situation-related object-state representations are activated, at which point in time do eye movements shift to the most plausible object-state? And conversely, at which moment in time does covert attention shift *away* from an implausible object state? To infer the object-state representation that participants had in mind, and the object-state representation that participants rejected as incompatible with the situation, we recorded their eye movements to the visual display (see Figure 1) while they listened to sentences via loudspeakers. Two competing states of the target object were depicted, an intact state (e.g. an intact onion) and a changed state (e.g., a chopped onion), together with two unrelated distractors.

While our primary interest is in the time-course of eye movements when participants were expected to shift their attention from one object state to the other, this study also affords the opportunity to address eye movements in anticipation of, and during, "*the onion*" in the first sentence, as a function of the verb ("*chop*" vs. "*weigh*"). Using the picture verification paradigm, Kang, Eerland, Joergensen, Zwaan, & Altmann (2019) found that participants reacted faster to a target probe object (e.g. a dropped ice cream cone) when it matched the end state of the event ("*The woman will drop the ice cream*") than when it (most likely) did not ("*The woman will choose the ice cream*"). However, using the VWP, Altmann & Kamide (2007) revealed that participants launched anticipatory eye movements more towards the object that afforded the action (e.g. a full glass of beer after '*the man will drink…*')

than at an object that depicted a plausible end state (e.g. an empty glass of wine). The current study will provide additional data to clarify this apparent discrepancy between the two sets of results.

In the present study, we examine whether participants might look at the object that affords the change (i.e. the intact onion in "*chop the onion*"), as might be predicted by Altmann & Kamide (2007). However, they might also, or instead, retrieve the object representation that matches the most probable end state as described in linguistic context (e.g. more looks at the intact state when a minimal change of state is described, and more looks at the chopped state when a substantial change is described), as might be predicted by Kang et al. (2019). Crucially, we might observe preferential looks to one or other depiction of the onion (intact or chopped), because that depiction is a 'good' fit for the most active mental representation, or because it is a 'good enough' fit for (i.e. overlaps sufficiently with) the most active representation. Having a single depiction in the visual scene would not allow us to distinguish between 'good' or 'good enough' fits, but having the two depictions allows us to establish which one is better ('good') than the other ('good enough'). We return to the significance of these two depictions in the Discussion section below.

Materials and Methods

Participants

64 students from the University of York participated in this study. They received either half an hour course credit or two pounds for their contribution. All participants were native speakers of British English and had normal or corrected to normal vision. Informed consent was obtained for the experiment from all participants. The protocol

of this study was approved by the Departmental Ethics Committee of the Department of Psychology at the University of York, UK.

Stimuli

Experimental stimuli consisted of 36 sets of sentences. The first sentence described Degree of Change (a minimal change vs. a substantial change) that is described to happen on the target object. The second sentence began with "but first/and then" that either indicated a backward or a forward shift of time. We will call this manipulation the Temporal Context (TC) from now on. Table 1 shows a set of example sentences and their corresponding experimental conditions. All sentences were recorded by a male native speaker of British English at a sampling rate of 44,100 Hz and 16-bit resolution in a soundproof booth. Identical sections in the auditory stimuli within the same set were cross-spliced and used across conditions. That is, in condition 1a) and 1b), only the Temporal Context was different in the recording, while in condition 1a) and 1c), everything else was from the same section of recordings other than the verbs. To ensure that the repeated noun phrase in the 2nd sentence was felicitous, the immediately preceding verb/preposition received contrastive stress ("And THEN / But FIRST, he will SMELL the onion" or "And THEN / But FIRST, he will walk AROUND the dustbin").

Table 1. Example sentences in Experiment 1.

Temporal Context (TC)	**Degree of Change (DC)**	
	Minimal ("weigh")	**Substantial ("chop")**
But first	**1a)** The chef will **weigh** the onion. **But first**, he will smell the onion.	**1c)** The chef will **chop** the onion. **But first**, he will smell the onion.
And then	**1b)** The chef will **weigh** the onion. **And then**, he will smell the onion.	**1d)** The chef will **chop** the onion. **And then**, he will smell the onion.

36 visual displays (see an example in Figure 1) were created using commercially available Clipart packages and presented on a 21” monitor at a resolution of 600x800 pixels. Each of the displays contained four concrete objects that were visually distinct from each other, including the target object in its intact state and changed state along with two unrelated distractors. The locations of the four depicted items were counterbalanced across experimental stimuli and the depicted objects were equally likely to occur in all four quadrants.

<<Insert Figure 1 here>>

Figure 1. An example visual stimulus used in Experiment 1. Participants heard “*The chef will* ***weigh/chop*** *the onion.* ***But first/And then****, he will smell the onion”.* The visual stimulus included an intact version and a changed version of the target object and two unrelated distractors.

A further 36 sentence/picture pairs were added as foils. These items had the same sentence structure as the experimental items. The visual displays for the foil items also showed four objects. In each display, there were two objects of the same category (e.g., two different-looking baskets) but these were not mentioned in the sentences; instead, one of the two other items were mentioned.

Experimental items were rotated across 4 lists in a Latin Square Design. In total, participants were presented with 36 experimental trials and 36 foils in each list.

Procedure

The experiment was conducted on an Eyelink II Head-mounted eye-tracker, which sampled at 500 Hz from the right eye, but viewing was binocular. Participants were seated in front of a 21" LED monitor with their eyes 60 cm from the display. They were instructed to inspect the visual display freely while listening to the sentences. Participants previewed the visual display for 1000ms before the sentences started playing over loudspeakers. Trials were presented in a fixed randomized order with 4 counter-balanced lists. Between each trial, a centered black dot was used to correct calibration drift. A nine-point calibration and validation procedure was performed after every six trials. Participants were informed when they had completed 50% of the trials and were encouraged to take a short break before completing the second half. Each trial lasted about 9 seconds and the total length of the experiment was about 25 minutes.

Data processing

Data were analyzed with the following procedure: First, four equally sized areas of interest (AOIs) were marked on each visual display, including the intact version of the target object, the changed version of the target object, and the two unrelated

distractors", as shown in Figure 2.

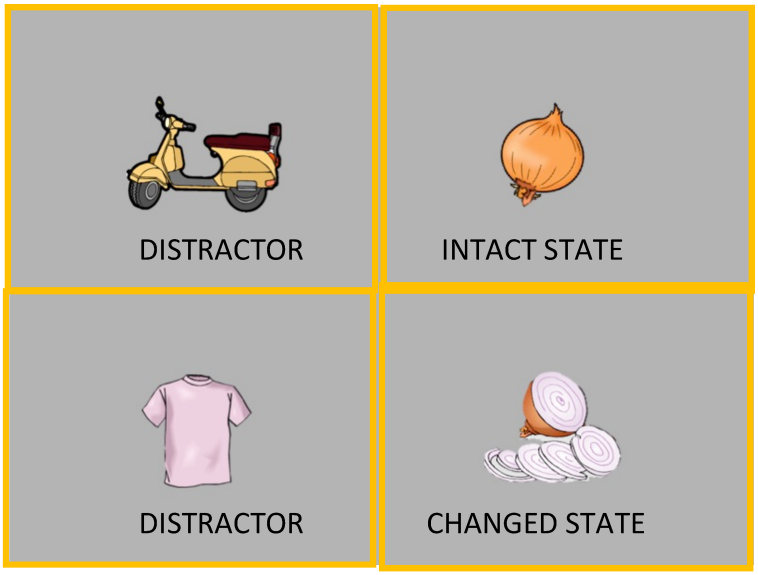


Figure 2. Experiment 1: Example of interest areas (AOIs) on a visual stimulus

Second, we marked the onset and offset of words in the sentences in their corresponding speech files using a commercial software known as "Sound Studio". We focused on three critical phrases of sentential stimuli in data analysis, including 1) the object name in the first sentence (e.g., "*the onion*"), 2) Temporal Context in the second sentence ("*But first/And then*"), and 3) the object name in the second sentence (e.g., "*the onion"*). These time windows were selected *a priori*.

We report two measures of eye movements – fixations and saccades – for each critical time window (see Altmann & Kamide, 2009 for discussion of the utility of these distinct measures). We report the probability of fixation at the onsets and offsets of our phrases of interest, to establish where attention was directed prior to the onset and after the presentation of the linguistic material that might cause shifts in that attention towards one depiction or another in the scene. Fixations at the onset of a region reflect anticipatory eye movements (as well as baseline fixations), and

fixations at the offset reflect shifts in attention between onset and offset. We report the probability of making one or more saccades during our phrases of interest towards the object-states, to establish the likelihood of a shift in attention to one depiction or another during the critical phrases (i.e. the time window between the onset and the offset of the critical phrases).

Our primary interests in the current experiment were twofold: First, to explore whether listeners display a bias in the first sentence to look towards the object-state that matches the described end state of the target object (as predicted by Kang et al., 2019), or whether they display a bias to look towards the initial state of the target object, corresponding to the state that would be acted upon in the first sentence (as predicted by Altmann & Kamide, 1999, 2007). Second, to establish the time course with which attention in the second sentence towards the depictions of the different target states is modulated by the interaction between Temporal Context (But first vs. And then) and Degree of Change described in the first sentence (minimal vs. substantial; e.g. "*chop the onion*" vs. "*weigh the onion*").

Statistical analyses on the data were conducted in R (R Core Team, 2019) using generalized linear mixed-effects models for binary dependent variables of the glmer function in the lme4 package (Bates et al., 2015). Each eye-tracking data sample was coded as a binomial outcome (depending on whether a participant looked at the AOI: 0= No Hit; 1 = Hit). Separate statistical analyses were conducted for the intact state and the change state. In the first sentence, Degree of Change (Substantial vs. Minimal) was included as a fixed effect with participants and items as random effects (Baayen et al., 2008). In the second sentence, Degree of Change (Substantial vs. Minimal) and Temporal Context (But first vs. And then) and their

interactions were included as fixed effects with participants and items as random effects. We assigned sum-coded contrasts to Degree of Change (Minimal = -1; Substantial = 1) and Temporal Context (And then = -1; But first = 1). We used non-maximal models without individual slopes for the random factors for our data because not all maximal models at each critical phrase converged. Please see an example model below:

Model <- glmer (Hit ~ Degree of Change* Temporal Context + (1 | participants) + (1 | items), family=binomial).

We report regression coefficients (β), standard errors (*SE*), and Wald's *Z*-score. We examined significance of the predictors via a likelihood ratio test between models with and without a fixed effect of the predictor. Post-hoc analysis after a significant interaction was conducted using linear contrasts with the emmeans function in the emmeans package (Lenth, 2018).

Results

Figure 3 show the percentage of trials with fixations on the two regions of interest (intact state vs. changed state of the target object), sampled in 25 ms increments from the onset of, and during, the presentation of sentences. As noted in Altmann and Kamide (2004, 2009), the unfolding speech and the unfolding eye movements are not synchronized in the graph because the graph displays average onsets and offsets of time windows across trials. Thus, the line graphs were resynchronized at the onset of each phrase to reflect the probability of fixation at the corresponding moments in time across all trials at the intersection between each curve and each vertical synchronization line. Table 2 presents percentage of trials with fixations on,

or saccades to, the two target AOIs and distractors, at the onset of, and during, at the offset the presentation three critical phrases in sentences. Results of the logistic regression models are presented in Table 3.

<<Insert Figure 3>>

<<Insert Table 2>>

<<Insert Table 3>>

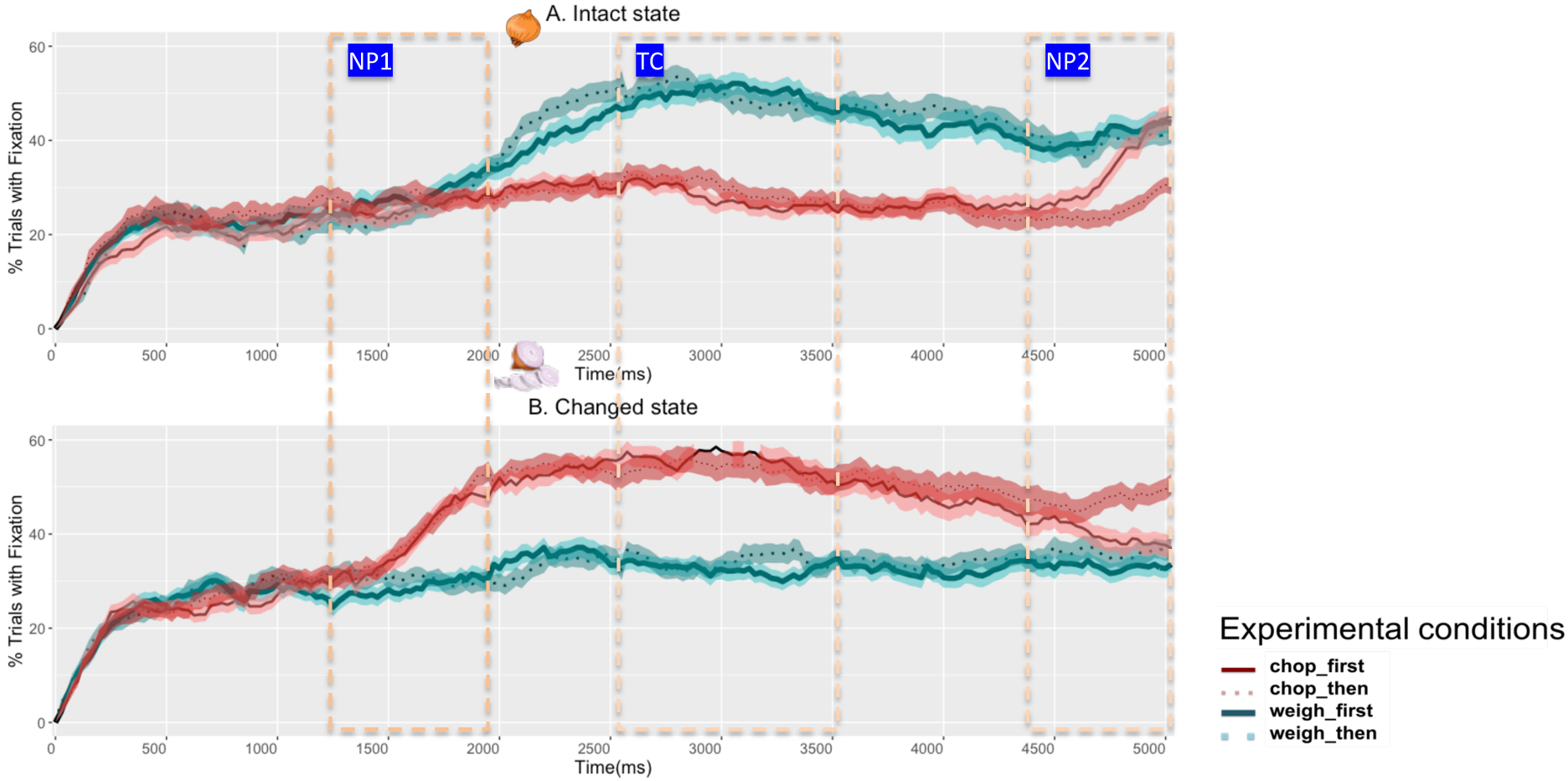


Figure 3. Experiment 1*:* Percentage of trials with eye movements launched on the interest areas (AOIs) across sentential conditions. The x-axis shows the elapsed time in successive 25 ms increments from the onset of sentential stimuli (e.g., "*The chef will weigh/chop* ***the onion*** *(NP1)****. But first/And then*** *(TC), he will smell* ***the onion*** *(NP2)").* The y-axis shows percentage of trials with at least one fixation on the AOIs. Standard errors above and below the mean were shown as shaded areas.

Table 2. Means and standard deviations (SD) for the percentage of trials with fixations and saccades at the three critical phrases in Experiment 1. Fixations were calculated on a trial-by-trial basis at the onset and at the offset of each critical phrase. Saccades were calculated on a trial-by-trial basis and resynchronized at the onset of each critical phrase.

		Intact state		Changed state		Distractors	
		Weigh	**Chop**	**Weigh**	**Chop**	**Weigh**	**Chop**
NP1	**Onset**	26 (15)	26 (15)	29 (14)	36 (17)	21 (16)	17 (13)
	Saccade	43 (17)	34 (17)	40 (18)	44 (16)	19 (14)	12 (13)
	Offset	42 (19)	30 (16)	35 (16)	53 (18)	10 (11)	7 (9)
TC	**Onset**	52 (20)	31 (16)	33 (16)	55 (20)	7 (10)	6 (10)
TC: But first	**Saccade**	26 (12)	22 (16)	27 (16)	28 (16)	12 (13)	13 (13)
	Offset	47 (20)	26 (15)	33 (16)	53 (23)	9 (11)	9 (11)
TC: And then	**Saccade**	22 (18)	19 (14)	24 (17)	22 (15)	10 (12)	12 (11)
	Offset	48 (20)	27 (17)	33 (18)	52 (22)	9 (11)	10 (10)
NP2: But first	**Onset**	39 (21)	28 (17)	34 (16)	41 (18)	12 (11)	12 (11)
	Saccade	32 (18)	39 (18)	31 (18)	24 (18)	14 (12)	14 (12)
	Offset	45 (20)	46 (19)	35 (14)	35 (18)	9 (11)	8 (10)
NP2: And then	**Onset**	36 (18)	23 (16)	37 (19)	45 (20)	15 (13)	15 (12)
	Saccade	29 (17)	30 (17)	26 (18)	31 (18)	11 (12)	12 (12)
	Offset	39 (21)	34 (16)	37 (17)	47 (19)	10 (10)	8 (10)

Notes: Weigh = minimal change (weigh the onion); Chop = substantial change (chop the onion); NP1 = noun phrase in the first sentence (the onion); TC = Temporal Context (but first/and then); NP2=noun phrase in the second sentence (the onion). Percentage of trials with eye movements on distractors was averaged across the two distractors on the same visual stimulus.

Table 3a. Experiment 1: Fixed effects and interactions from the mixed-effects logistic regression models on the probability of looking at the **intact state**. Fixations were calculated on a trial-by-trial basis at the onset and the offset of the critical phrases. Saccades were calculated on a trial-by-trial basis and resynchronized at the onset of each critical phrase.

		NP1			TC			NP2		
		Onset	**Saccade**	**Offset**	**Onset**	**Saccade**	**Offset**	**Onset**	**Saccade**	**Offset**
(Intercept)	***β***	-1.06	-0.49	-0.61	-0.37	-1.29	-0.57	-0.80	-0.75	-0.36
	SE	0.06	0.09	0.07	0.07	0.08	0.06	0.06	0.08	0.06
	Z	-17.80***	-5.32***	-9.08***	-5.47***	-16.02***	-9.55***	-12.69***	-9.65***	-5.79***
Degree of Change (DC)	***β***	0.01	-0.21	-0.29	-0.47	-0.12	-0.48	-0.31	0.09	-0.06
	SE	0.06	0.04	0.07	0.06	0.05	0.06	0.06	0.05	0.06
	Z	0.17	-4.62***	-4.32***	-7.43***	-2.35*	-8.40***	-5.35**	2.06*	-1.01
Temporal Context (TC)	***β***					0.10	-0.02	0.10	0.13	0.19
	SE					0.05	0.06	0.06	0.05	0.06
	Z					2.01*	-0.42	1.71	2.86**	3.52***
Degree of Change x Temporal Context (DC x TC)	***β***					-0.01	-0.01	0.03	0.07	0.07
	SE					0.05	0.06	0.06	0.05	0.06
	Z					-0.14	-0.05	0.61	1.53	1.29

Notes*: DC = Degree of Change; NP1 = object name in the first sentence (e.g., "the onion"); TC = Temporal Context (i.e. "but first/and then"); NP2 = object name in the second sentence (e.g., "the onion");* *$p < .05$; ** $p < .01$; *** $p < .001$

Table 3b. Experiment 1: Fixed effects and interactions from the mixed-effects logistic regression models on the probability of looking at the **changed state**. Fixations were calculated on a trial-by-trial basis at the onset and the offset of the critical phrases. Saccades were calculated on a trial-by-trial basis and resynchronized at the onset of each critical phrase.

		NP1			TC			NP2		
		Onset	**Saccade**	**Offset**	**Onset**	**Saccade**	**Offset**	**Fixation**	**Saccade**	**Offset**
(Intercept)	***β***	-0.75	-0.35	-0.25	-0.27	-1.12	-0.30	-0.45	-1.02	-0.47
	SE	0.06	0.10	0.07	0.08	0.08	0.07	0.08	0.11	0.06
	Z	-13.15***	-3.44***	-3.68***	-3.52***	-14.19***	-4.21***	-6.00***	-9.67***	-7.83***
Degree of Change (DC)	***β***	0.18	-0.09	0.42	0.50	-0.01	0.43	0.17	-0.03	-0.11
	SE	0.06	0.04	0.06	0.06	0.05	0.05	0.06	0.05	0.05
	Z	3.13**	2.11*	6.50***	8.02***	-0.28	8.45***	2.73**	-0.61	2.10*
Temporal Context (TC)	***β***					0.13	0.02	-0.09	-0.02	-0.15
	SE					0.05	0.05	0.06	0.05	0.05
	Z					2.57*	0.32	-1.41	-0.46	-2.88**
Degree of Change x Temporal Context (DC x TC)	***β***					0.05	0.01	0.02	-0.15	-0.10
	SE					0.05	0.05	0.06	0.05	0.05
	Z					0.94	0.25	-0.33	-3.21**	-1.88

Notes*: DC = Degree of Change; NP1 = object name in the first sentence (e.g., "the onion"); TC = Temporal Context (i.e. "but first/and then"); NP2 = object name in the second sentence (e.g., "the onion");* **p* < .05; ** *p* <.01; *** *p* <.001

First critical noun phrase: Object name in the first sentence (NP1)

To examine the anticipatory influence of the verb ("*weigh*" vs. "*chop*") prior to the onset of the object name, we examined proportions of fixations on the intact and changed states of the onion at the onset of the phrase "*the onion*". There was no difference in the probability of fixating the intact state across the "*chop*" and "*weigh*" conditions (χ^2 = 0.03, p = .865). By contrast, participants were already fixating, at the onset of the critical noun phrase, the changed state more in the "*chop*" condition than in the "*weigh*" condition (χ^2 = 9.55, p = .002). Thus, whereas anticipatory looks towards the chopped onion were modulated by the verb, that was not the case for looks towards the intact onion.

During "*the onion*", there was a significant effect of Degree of Change (χ^2 = 21.42, p < .001) with a higher probability of launching a saccadic eye movement towards the intact state in the "*weigh*" condition than the "*chop*" condition. Conversely, the probability of launching a saccadic eye movement towards the changed state was higher in the "*chop*" condition than the "*weigh*" condition (χ^2 = 4.43, p = .035). Note that on any given trial, participants could in principle launch one *or more* saccades towards the target depiction (if, for example, they moved to it, then moved away and then moved back again) – the calculation of probability treats one or more saccades as a single binary outcome on a given trial ("none" or "one or more").

At the offset of the phrase "*the onion*", the probability of fixating the intact state was higher in the "*weigh*" condition than the "*chop*" condition (χ^2 = 17.76, p < .001), while the changed state showed the opposite pattern (χ^2 = 37.82, p < .001), suggesting higher probability of fixating the end state of the object as inferred from the language context.

Thus, results from the first sentence suggest that, after an action was referred to that entailed a substantial change to the object acted upon, visual attention was directed more towards the changed visual depiction of the object (e.g., a chopped onion) when it matched the end state inferred from the language context ("*chop*") than when it mismatched (e.g., "*weigh*"). This difference in visual attention to the changed state emerged before the target object was named. No such difference in anticipatory eye movements was found for the intact state (e.g., a whole onion).

Second critical phrase: Temporal Context (TC)

At the onset of "*but first/and then*", the probability of fixating the intact state was higher in the "*weigh*" condition than in the "*chop*" condition ($\chi^2 = 48.58$, $p < .001$). Conversely, participants were more likely to be looking at the changed state in the "*chop*" condition than in the "*weigh*" condition ($\chi^2 = 54.93$, $p < .001$). These patterns reflect the impact of the first sentence on the fixation record of both object states at the onset of the second sentence.

During "*but first/and then*", we observed significant effects of Temporal Context on saccades to the intact state ($\chi^2 = 4.07$, $p = .044$) and to the changed state ($\chi^2 = 6.56$, $p = .010$): Participants were more likely to saccade towards both the *intact and* changed states in the "*but first*" condition than in the "*and then*" condition. We also found a significant effect of Degree of Change that participants were more likely to saccade towards the intact state in the "*weigh*" condition compared to the "*chop*" condition ($\chi^2 = 5.54$, $p = .019$). However, no effect of Degree of Change was found on the changed state ($\chi^2 = 0.05$, $p = .818$). No interactions between Degree of Change and Temporal Context were found (Intact state: $\chi^2 = 0.02$, $p = 0.888$; Change state: $\chi^2 = 0.89$, $p = .347$).

At the offset of "*but first/and then*", there were neither significant effects of

Temporal Context on the intact state (χ^2 = 0.18, p = .674) and the changed state (χ^2 = 0.11, p = .740), nor any interactions (Intact state: χ^2 = 0.002, p = .957; Changed state: χ^2 = 0.06, p = .806). However, there were still significant effects of Degree of Change on both the intact state (χ^2 = 60.11, p < .001) and the changed state showed the opposite pattern (χ^2 = 60.69, p < .001) that there was a higher probability of fixating intact state in the "*weigh*" condition than the "*chop*" condition with the changed state in the opposite pattern.

The results thus suggest that both prior to and after "*but first/and then*", participants' fixations were still primarily influenced by the linguistic context of the first sentence, despite the higher probability of launching saccadic eye movements to both the intact and the changed states in the "*but first"* condition than the "*and then*" condition.

Third critical phrase: Object name in the second sentence (NP2)

At the onset of "*the onion*" in the second sentence, there were no significant effects of Temporal Context (χ^2 = 2.77, p = .100) and no interaction between Degree of Change and Temporal Context (χ^2 = 0.37, p = .542) on the probability of fixating the intact state. There was a significant effect of Degree of Change: the intact state was fixated more in the "*weigh*" condition than in the "*chop*" condition (χ^2 = 26.81, p < .001). Similarly, there were no significant effects of Temporal Context (χ^2 = 2.00, p = .157) on the probability of fixating the changed state, and no interaction with Degree of Change (χ^2 = 0.11, p = .741) but there was a significant effect of Degree of Change (χ^2 = 7.33, p = .007) with a higher probability of fixating the changed state in the "*chop*" condition than the "*weigh*" condition regardless of temporal context.

During "*the onion*", there *was* a significant effect of Temporal Context on the probability of making saccadic eye movements towards the intact state (χ^2 = 8.47, p

= .004) with a higher probability in the "*but first*" condition than the "*and then*" condition. There was also a significant effect of Degree of Change (χ^2 = 4.54, p = .033), with more saccadic eye movements launched towards the intact state in the "*chop*" condition than the "*weigh*" condition. This pattern was opposite to what was observed in previous critical time windows – there, looks were more likely to be directed towards the intact state in the "*weigh*" condition than in the "*chop*" condition (reflecting impact from the first sentence). There was no interaction between Degree of Change and Temporal Context (χ^2 = 2.33, p = .127). Inspection of Table 2 suggests that participants' visual attention towards the intact state was driven by the influence of "*but first*". In contrast, there was a significant interaction between Degree of Change and Temporal Context on the probability of saccadic eye movements towards the changed state (χ^2 = 10.25, p = .001). Post-hoc analysis revealed that participants were more likely to look towards the situationally-appropriate changed state in "*chop, and then*" than in the "*chop, but first*" condition (Odds Ratio (OR) = 1.38, 95% Confidence Interval (CI): 1.06, 1.79). Thus, the results suggested that participants moved their eyes towards the situationally appropriate object-state as determined by the temporal context (and the action described in the first sentence).

At the offset of "*the onion*" in the second sentence, there was no significant effect of the Degree of Change on the probability of fixating the intact state (χ^2 = 0.95, p = .330), but there was a significant effect of Temporal Context (χ^2 = 11.85, p < .001): Participants were more likely to fixate the intact state in the "*but first*" condition than the "*and then*" condition. There was no interaction between Temporal Context and Degree of Change on the probability of fixating the intact state (χ^2 = 1.65, p = .199). Similarly, for fixations on the changed state: There was no significant effect of Degree of Change (χ^2 = 4.40, p = .359), but a significant effect of Temporal

Context (χ^2 = 8.00, p = .005) – participants were more likely to fixate the changed state in the "*and then*" condition than in the "*but first*" condition, which was opposite to the pattern of the intact state. However, the interaction between Temporal Context and Degree of Change only approached significance (χ^2 = 3.51, p = .061).

The results thus suggest that upon hearing the noun phrase in the second sentence, participants continued to adjust their visual attention to the intended end state of the target object. At the offset of the noun phrase, participants were more likely to be fixating the situationally-appropriate object-state.

Discussion

Experiment 1 demonstrated that language guides eye movements towards depictions of the situationally-appropriate states of the target object. When hearing that an onion will be chopped "*… but first* ..." the eyes move towards the depiction of a chopped onion as the description unfolds, but then end up on the situationally appropriate depiction of an intact (i.e. pre-chopped) onion. And when hearing that an onion will be chopped *"... and then* ..." the eyes tend to remain on the depiction of the chopped onion. Perhaps surprisingly, in the "*chop... but first*..." condition, we did not observe *anticipatory* eye movements towards the intact onion – i.e. before the object was referred to directly until at the end of the second sentence. In interpreting these data (see the General Discussion, below), we make a number of assumptions about the interpretation of a sentence or discourse in the context of a concurrent (or previous) visual scene. First, we assume, following Altmann & Kamide (2007) that a sentence is a dynamically unfolding representation of an *event*, and that eye movements towards or away from objects in a scene reflect real-time changes in the similarity between the dynamically changing mental event structure and the objects in the scene. Second, we assume that looks from one visual instantiation of the

onion (e.g. the chopped onion) to another (e.g. the intact onion) do not necessarily indicate a “commitment” to those two visual depictions referring to the *same* individual token onion as referenced by the language (in the real world, the intact and chopped “versions” of the token onion cannot co-exist in time). Rather, we assume that looks to one or other depiction reflect featural overlap between one or other visual object and the features activated as a part of the dynamically updating event structure being constructed as the concurrent language unfolds.

By limiting the visual depictions to the initial and end states of the chopping, we necessarily coerce an underlying continuous and dynamical representation of the onion – changing through time from intact to chopped – into two discrete instances at each end of the transformational continuum (see Altmann & Ekves, 2019, for a theoretical account of dynamical object representations as trajectories through feature space across time). Nonetheless, the utility of the paradigm, notwithstanding this discontinuity of visual representation, lies in its ability to use eye movements to one or other objects in the scene as a way of probing the changing mental representations activated during language comprehension: By displaying two versions of the same object, we were able to probe the object-representation that listeners constructed and modified as the language unfolded. What we observed in the “*chop ... but first* ...” cases is that the closest fit between the visual depiction of the intact onion and the (mental) representational correlates of the unfolding sentence occurred when “*the onion*” was referenced at the end of the second sentence, while before this point, there was a closer fit between these unfolding representational correlates and the visual depiction of the *chopped* onion.

We return to discussion of this late effect of Temporal Context (and the patterns we observed during the first sentence) in the General Discussion. But first,

we turn to a conceptual replication, using a different sentence type.

Experiment 2

In Experiment 2, we sought to explore whether the above time-course effects would replicate if the change of state were manipulated in a different way. In this Experiment, Degree of Change (a minimal change vs. a substantial change) was manipulated by using two different objects in the first sentence (*"The girl will stomp on the penny/egg"*; c.f. Hindy et al., 2012). In the condition corresponding to *'stomp on the penny'* there is, most likely, no corresponding change of state (in the real-world correlate to this event description), but in '*stomp on the egg*' we do expect such a change. Like in Experiment 1, the second sentence in Experiment 2 referred back to the object introduced in the first sentence (*"But first/And then, she will look down upon the penny/egg"*).

Materials and Methods

Participants

64 students from the University of York participated in this study. None took part in Experiment 1. They were all native speakers of British English with normal or corrected to normal vision. They received either two pounds or half an hour course credit for their participation. Informed consent was obtained for the experiment from all participants. The protocol of this study was also approved by the Departmental Ethics Committee of the Department of Psychology at the University of York, UK.

Materials

Linguistic stimuli. 36 sets of auditory stimuli were used. In each stimulus, the first sentence described either a minimal change or a substantial change that involved the same action with two different objects (e.g., "*stomp on the egg vs. the penny*). The

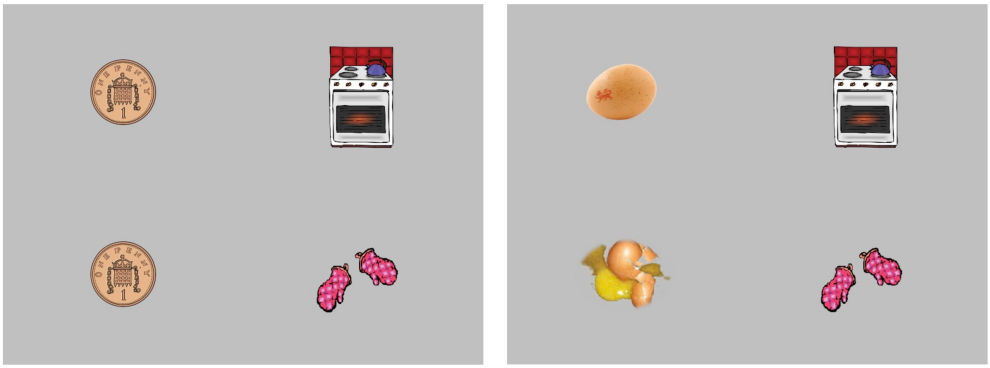
second sentence indicated a backward ("*but first*") or a forward ("and then") shift of time. See a set of example sentences in Table 4.

Table 4. Example sentences in Experiment 2.

Temporal Context (TC)	Degree of Change (DC)	
	Minimal ("penny")	**Substantial ("egg")**
But first	2a) The girl will stomp on **the penny**. **But first**, she will look down at **the penny**.	2c) The girl will stomp on **the egg**. **But first**, she will look down at **the eg**g.
And then	2b) The girl will **stomp on the penny**. And then, she will look down at the penny.	2d) The girl will stomp on **the egg**. **And then**, she will look down at **the egg**.

Visual stimuli. Each set of stimuli was accompanied by a pair of visual displays that were created using commercially available Clipart packages (see an example in Figure 4). Whereas for the egg example (2c and 2d) the scene depicted one egg in its intact state and one egg in a changed state, for the penny examples (2a and 2b) the scene depicted two identical versions of the penny. Our reasoning behind two identical versions follows our earlier logic (from Altmann & Kamide, 2007) that looks towards an object in the visual scene during concurrent language reflects the goodness of fit between the semantic (including form) features activated by the visual depictions, and the semantic features activated by that unfolding language. We thus anticipated that looks would be split equally between the two pennies when "*the penny*" was referred to. They thus provide a baseline against which to interpret looks to the different depictions of the egg. Counterbalancing the location of each object aimed to eliminate the biases that one often sees in eye movement studies that often favor looks towards one quadrant at the expense of looks towards others. 36 foil items that were the same as in Experiment 1 were included.

<<Insert Figure 4 below>>



"the penny" "the egg"

Figure 4. Experiment 2: Example of an experimental visual stimulus. The corresponding auditory stimuli were *"The girl will stomp on the penny/egg. But first/And then, she will look down at the penny/egg".*

Procedure

The procedure was the same as Experiment 1.

Results

Data were analyzed in the same way as Experiment 1. Again, for the sake of exposition, we refer to the target objects with the example "*The girl will stomp on the penny/egg. And then/But first, she will look down at the penny/egg"*. We report statistical results for "*the penny*" and "*the egg*" separately, so we can compare whether there were differences in eye movements in the two conflicting versions of "the egg" but two identical versions of "the penny". We sum-coded contrasts of the predictors (State: Intact state = -1; Changed state = 1; Temporal Context: And then = -1; But first = 1). We report regression coefficients (β), standard errors (*SE*), and Wald's Z-score. We also report significance of the predictors via a likelihood ratio test between models with and without a fixed effect of the predictor. Post-hoc analysis after a significant interaction was conducted using linear contrasts with the emmeans function in the emmeans package (Lenth, 2018). Please see below for an example model:

Model <- glmer (Hit ~ State* Temporal Context+ (1 | participants) + (1 | items), family=binomial).

Critical phrases include verbs in the first sentence (e.g., "*stomp on*"), the first reference to the target objects (e.g., "*the penny/egg*"), Temporal Context ("*but first/and then*"), and the second reference to the target object (e.g., "*the penny/egg"*). The two states of the penny in the accompanying visual display were identical, but were labeled as "*intact*" and "*changed*" as determined by the locations of the corresponding states of the egg (so whichever penny occupied the location of the intact egg in the corresponding display was labeled "*intact*").

Figure 5 presents the percentage of trials with fixations on target objects in 25 ms increments from the onset of, and during the presentation of auditory stimuli in Experiment 2. Table 5 presents the percentage of trials with fixations on, and saccades towards, the two target objects and distractors at the onsets of, during, and at the offsets of the presentation of the three critical phrases. Table 6 presents results of the logistic regression models in Experiment 2.

<<Insert Figure 5>>

<<Insert Table 5>>

<<Insert Table 6>>

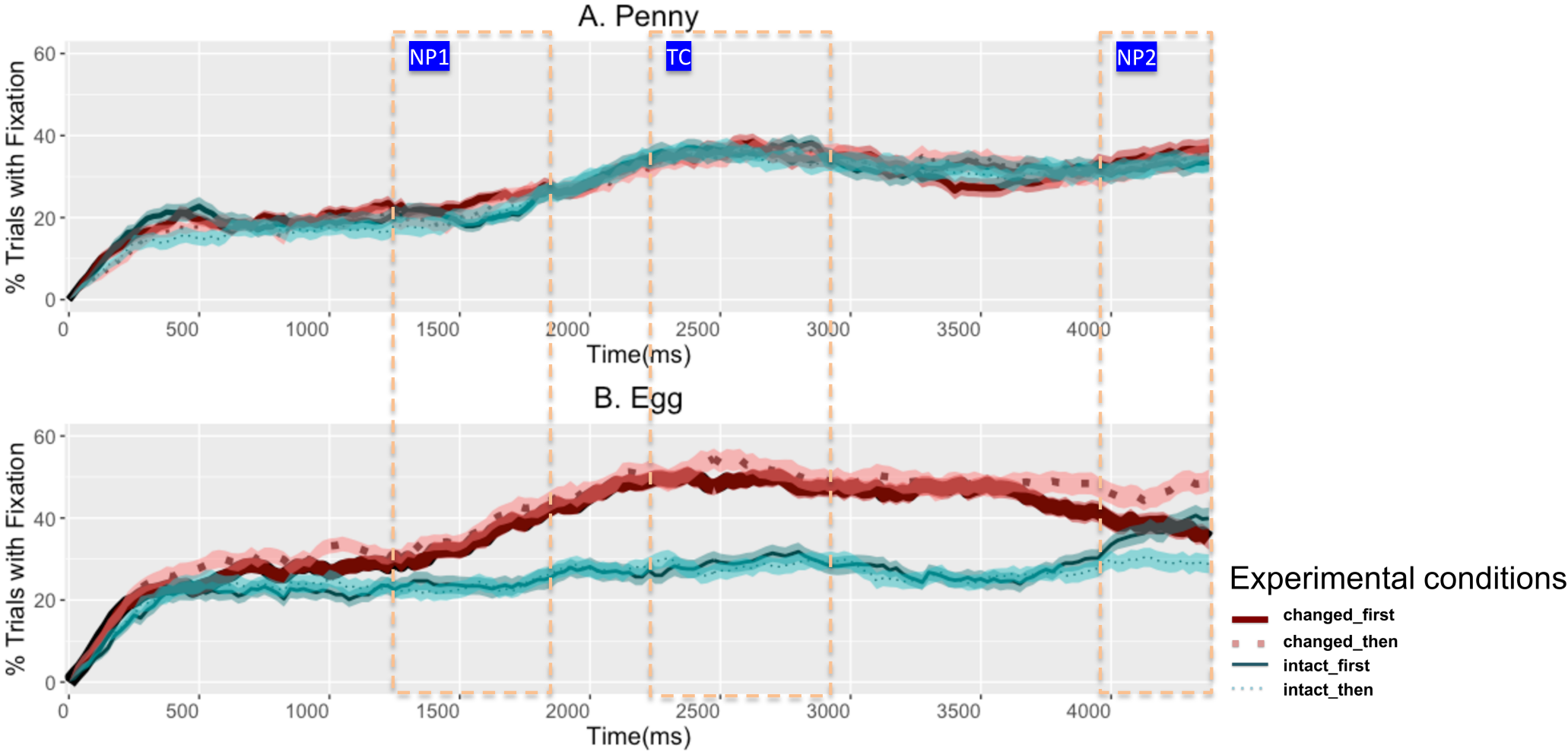


Figure 5. Experiment 2*:* Mean percentage and standard deviations of trials with fixations on the AOIs as sentences unfolded. The x-axis shows the elapsed time in successive 25 ms time windows from the onset of sentential stimuli (e.g., "*The girl will stomp on the penny/egg. But first/And then, she will look down at the penny/egg")*. The y-axis shows the percentage of trials with at least one fixation. Standard errors above and below the mean were shown as shaded areas. The two depicted states of the object in the minimal change condition (e.g., penny) were exactly the same. Thus, labels "changed/intact" of the 'penny' condition refer to the states of the 'egg' in the corresponding locations.

Table 5. Means and standard deviations (SD) for the percentage of trials with fixations and saccades during the three critical phrases in Experiment 2. Fixations were calculated on a trial-by-trial basis at the onset and the offset of the critical phrases. Saccades were calculated on a trial-by-trial basis and resynchronized at the onset of each critical phrase.

		Penny			Egg		
		'Intact'	**'Changed'**	**Distractors**	**Intact**	**Changed**	**Distractors**
NP1	**Onset**	18 (14)	18 (13)	24 (15)	20 (14)	30 (14)	17 (14)
	Saccade	29 (15)	29 (15)	20 (13)	29 (16)	37 (15)	15 (13)
	Offset	28 (15)	27 (14)	15 (11)	24 (14)	41 (15)	9 (10)
TC	**Onset**	31 (14)	32 (15)	10 (10)	24 (15)	42 (16)	6 (9)
TC: But first	**Saccade**	28 (17)	28 (16)	17 (14)	25 (14)	27 (14)	12 (12)
	Offset	28 (16)	30 (16)	13 (9)	25 (14)	40 (17)	9 (11)
TC: And then	**Saccade**	26 (14)	27 (14)	14 (14)	21 (12)	21 (12)	9 (11)
	Offset	29 (14)	28 (17)	12 (13)	22 (14)	42 (16)	7 (10)
NP2: But first	**Onset**	28 (16)	30 (18)	13 (14)	30 (15)	35 (17)	9 (11)
	Saccade	18 (13)	22 (13)	10 (11)	25 (12)	18 (12)	10 (10)
	Offset	27 (16)	34 (15)	11 (11)	34 (17)	29 (16)	8 (10)
NP2: And then	**Onset**	30 (15)	28 (15)	15 (12)	26 (15)	36 (17)	9 (10)
	Saccade	22 (14)	21 (14)	11 (11)	19 (11)	22 (14)	10 (12)
	Offset	32 (15)	30 (15)	11 (11)	26 (12)	39 (16)	8 (10)

Notes: Penny = minimal change (stomp on the penny); Egg = substantial change (stomp the egg); NP1 = noun phrase in the first sentence (the penny/egg); TC = Temporal Context (but first/and then); NP2 = noun phrase in the second sentence (the penny/egg). Percentage of trials with eye movements on distractors was averaged across the two distractors on the same visual stimulus. The two depicted states of the object in the minimal change (e.g., penny) were exactly the same. Thus, labels "changed/intact" of the 'penny' refer to the states of the 'egg' in the corresponding locations.

Table 6a. Experiment 2: Fixed effects and interactions from the mixed-effects logistic regression models on the probability of looking at the two identical objects in the minimal change condition (e.g., **two coins)**. Fixations were calculated on a trial-by-trial basis at the onset and the offset of the critical phrases. Saccades were calculated on a trial-by-trial basis and resynchronized at the onset of each critical phrase.

		NP1			TC			NP2		
		Onset	**Saccade**	**Offset**	**Onset**	**Saccade**	**Offset**	**Onset**	**Saccade**	**Offset**
(Intercept)	***β***	-1.40	-0.78	-0.87	-0.64	-0.88	-0.78	-0.79	-1.24	-0.69
	SE	0.05	0.07	0.05	0.05	0.06	0.05	0.05	0.06	0.05
	Z	-25.62***	-11.40***	-18.26***	14.04***	-14.66***	-15.86***	-15.04***	-19.99***	-14.87***
State	***β***	0.01	0.01	-0.02	-0.01	0.03	0.01	0.01	0.04	0.06
	SE	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	Z	0.10	0.05	-0.44	0.26	0.58	0.02	0.16	0.80	1.28
Temporal Context (TC)	***β***					0.04	0.01	0.02	-0.06	-0.02
	SE					0.05	0.05	0.05	0.06	0.05
	Z					0.81	0.32	0.35	-0.96	-0.43
State x Temporal Context (State x TC)	***β***					0.01	0.05	0.04	0.07	0.11
	SE					0.05	0.05	0.05	0.05	0.05
	Z					-0.22	0.96	0.83	1.39	2.47*

Notes*: NP1 = object name in the first sentence (e.g., "the penny/egg"); TC = Temporal Context ("but first/and then"); NP2 = object name in the second sentence (e.g., "the penny/egg").* The two depicted states of the object in the minimal change (e.g., penny) were exactly the same. Thus, labels "changed/intact" of the 'penny' refer to the same locations corresponding to the 'egg' condition. *p < .05; ** p <.01; *** p <.001.

Table 6b. Experiment 2: Fixed effects and interactions from the mixed-effects logistic regression models on the probability of looking at the two contrasting object states in the substantial change condition (e.g., **two eggs**). Fixations were calculated on a trial-by-trial basis at the onset and the offset of the critical phrases. Saccades were calculated on a trial-by-trial basis and resynchronized at the onset of each critical phrase.

		NP1			TC			NP2		
		Onset	**Saccade**	**Offset**	**Onset**	**Saccade**	**Offset**	**Onset**	**Saccade**	**Offset**
(Intercept)	***β***	-0.97	-0.58	-0.60	-0.55	-1.06	-0.60	-0.63	-1.22	-0.61
	SE	0.05	0.07	0.05	0.05	0.06	0.05	0.05	0.07	0.05
	Z	-19.64***	-8.76***	-12.85***	-11.86***	-17.70***	-12.88***	-13.67***	-17.75***	-13.22***
State	***β***	0.25	0.19	0.39	0.42	0.02	0.42	0.19	-0.05	0.08
	SE	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	Z	5.10***	4.12***	8.42***	8.98***	0.38	8.96***	4.01***	-0.10	1.81
Temporal Context (TC)	***β***					0.15	0.02	0.03	0.03	0.01
	SE					0.06	0.05	0.05	0.06	0.95
	Z					2.59**	0.40	0.59	0.49	-0.002
State x Temporal Context (State x TC)	***β***					0.04	-0.05	-0.04	-0.15	-0.22
	SE					0.05	0.05	0.05	0.05	0.05
	Z					0.86	-1.00	-0.86	-2.79**	-4.73***

Notes*: NP1 = object name in the first sentence (e.g., "the penny/egg"); TC = Temporal Context ("but first/and then"); NP2 = object name in the second sentence (e.g., "the penny/egg").* The two depicted states of the object in the minimal change (e.g., penny) were exactly the same. Thus, labels "changed/intact" of the 'penny' refer to the same locations corresponding to the 'egg' condition. *p < .05; ** p <.01; *** p <.001.

First critical phrase: Object name in the first sentence (NP1)

For "*the penny*", we did not find a significant effect of State (i.e. location in this condition) on the probability of fixations at the onset (χ^2 = 0.01, p = .922) or offset (χ^2 = 0.20, p = .658) nor the probability of saccades during "*the penny*".

By comparison, at the onset of NP1, participants were already more likely to fixate the broken egg than the intact egg (χ^2 = 26.36, p < .001). During NP1, participants were more likely to launch a saccade towards the broken egg than the intact egg (χ^2 = 17.09, p < .001), and at the offset of NP1, there was a higher probability of fixating the broken egg than the intact egg (χ^2 = 72.63, p < .001).

The results of the first sentence were thus consistent with Experiment 1. When a substantial change was expected from the action, participants tended to look more at a likely end state (i.e. a broken egg) than an initial state (i.e. an intact egg), even before the object's corresponding referring expression ("*the egg*").

Second critical phrase: Temporal Context (TC)

Unsurprisingly, no difference was found between looks towards the two pennies. Participants looked equally at the two versions of "the penny" at the onset (χ^2 = 0.07, p = .796), during (Interaction: χ^2 = 0.05, p = .825; TC: χ^2 = 0.65, p = .420; State: χ^2 = 0.01, p = .976), and at the offset (Interaction: χ^2 = 0.93, p = .336; TC: χ^2 = 0.10, p = .752; State: χ^2 = 0.34, p = .563) of hearing "*but first/and then*".

At the onset of "*but first/and then*", the probability of fixating the broken egg was still higher than the intact egg (χ^2 = 82.84, p < .001), suggesting the continuing impact of the first sentence on visual attention. During "*but first/and*

then", there was a significant effect of Temporal Context (χ^2 = 6.24, p = .011): Participants were more likely to launch saccadic eye movements in the "*but first*" condition than in the "*and then*" condition. There was no significant effect of State (χ^2 = 0.19, p = .662) nor any interaction (χ^2 = 0.74, p = .388) – in other words, eye movements were launched more in the "*but first*" condition regardless of whether towards the intact or broken egg. At the offset of "*but first/and then*", there was still no significant effect of Temporal Context (χ^2 = 0.07, p = .786), but there was a significant effect of State (χ^2 = 82.43, p < .001) with a higher probability of fixating the broken egg than the intact egg. No interaction between State and Temporal Context was found (χ^2 = 1.02, p = .314). Thus, consistent with Experiment 1, looks at the onset and offset of the temporal connective were driven primarily by continuing impact from the preceding sentence, notwithstanding the higher probability of launching saccades in the "*but first*" condition than the "*and then*" condition for "*the egg*".

Third critical phrase: Object name in the second sentence (NP2)

At the onset of NP2, there were again no significant effects of Temporal Context (χ^2 = 0.13, p = .722), State (χ^2 = 0.03, p = .865), nor their interaction (χ^2 = 0.68, p = .408) on the probability of fixating "the penny". During NP2, there were no differences in the probability of saccades towards "the penny" either. Interestingly, at the offset of NP2, however, there was a significant interaction between Temporal Context and State on the probability of fixating "the penny" (χ^2 = 6.13, p = .013). Post-hoc analysis showed that participants were more likely to fixate one penny than the other in the "*but first*" condition (OR = 1.41, 95% CI: 1.09, 1.83). We note that this bias appears to go in the opposite direction to the pattern of looks towards the corresponding locations in the 'egg'

conditions (see Table 5) and hence is likely to be spurious rather than a systematic bias due to e.g. location.

As for “the egg”, the pattern of fixations was similar to that observed at the offset of TC. There was no significant effect of Temporal Context (χ^2 = 0.29, p = .593) but there was a significant effect of State (χ^2 = 16.12, p < .001): Participants fixated the broken egg more than the intact egg. There was no interaction between the two (χ^2 = 0.73, p = .393). During NP2, there was an interaction between Temporal Context and State on the probability of saccades (χ^2 = 7.79, p = .005) towards the egg. Post-hoc analysis showed that participants were less likely to launch saccades towards the broken egg in the “*but first*” condition than in the “*and then*” condition (OR = 0.68, 95% CI: 0.51, 0.90). At the offset of NP2, there was again a significant interaction between Temporal Context and State (χ^2 = 22.54, p < .001). Post-hoc analysis suggested that participants were more likely to fixate the broken egg in the “*and then*” condition than in the “*but first*” condition (OR = 1.55, 95% CI: 1.20, 1.99), and more than the intact egg in the “*and then*” condition (OR = 1.83, 95% CI: 1.41, 2.37). Besides, participants were less likely to fixate the intact egg in the “*but first*” condition than “and then” condition (OR = 0.65, 95% CI: 0.50, 0.84).

The results thus suggested that in the substantial change condition (e.g., “the egg”), participants modified their visual attention towards the expected end state during the first sentence (and away from the intact/initial state), but during the second sentence after “*but first*” visual attention only shifted back towards the intact/initial state after the object was named again but not before (there was no anticipatory shift in eye movements towards the situationally appropriate, intact, egg after “*but first*”).

Discussion

In Experiment 2, Degree of Change was manipulated in the first sentence by using the same verb ("*stomp on"*) but two different nouns ("*the penny/egg"*). Separate visual displays were used depending on which kind of object was referred to. Two different object states (intact vs. changed) were depicted for the object that could change its state (e.g., "*stomp on the egg"*), but two identical versions were presented for the object that was not expected to change its state (e.g., "*stomp on the penny"*). The second sentence, like in Experiment 1, started with a Temporal Context that either indicated a backward ("*but first*") or a forward time shift (*"and then*"), further modulating the intended end state of the target object.

The results replicated Experiment 1: the linguistic context guided eye movements towards the situationally-appropriate end state of the target object, but it did so late – that is, at the point at which the object was referred to at the end of the sentence. Like in Experiment 1, and as is clear from Figure 4b, no difference in visual attention towards the intact egg was revealed between the "*but first*" condition and the "*and then*" condition (where it corresponds to the situationally *inappropriate* state) until after the onset of "*the egg*".

General discussion

Our experiments explored the time-course of retrieving object-state representations during (and following) the description of events in which objects changed state. Participants viewed visual displays that included two different versions of a target object, corresponding to the distinct physical states of the objects before and after an action that, depending on the experimental condition,

was described in a preceding sentence (e.g., an intact onion as the initial state and a chopped onion as the end state of a chopping action). The sentences that participants heard while viewing these displays described actions that either did or did not change the state of the target object; the changes in state were either conveyed by the verb (e.g., *weigh/chop the onion*), or could be inferred on the basis of the type of object being acted upon (e.g. *stomp on the penny/egg*).

Both studies yielded very similar patterns of eye movements: At the onset of the first mention of the target object in the first sentence, more looks were directed, or tended to be directed, towards the changed version of the target objects (i.e. the chopped onion in Experiment 1, and the broken egg in Experiment 2). This was not a non-linguistic bias to simply look more at a depiction of an object in a non-canonical, atypical, changed state: there were significantly more looks to the chopped onion and the broken egg after “chop” and “stomp on” than after “weigh” and “look down at” respectively. Similarly, *during* the first mention of the target object in the first sentence, more looks were directed towards the changed version of each target in the condition in which the language described a substantial change (Experiment 1) or in which a substantial change could be inferred (Experiment 2). Again, the eye movements *during* the critical noun phrase in the first sentence were not driven by typicality effects – it is certainly the case that, in many instances, the changed version of each object was the less typical, and as such may have attracted attention because being less typical entails being more informative. But the fact that attention was modulated by the verb suggests that typicality

was not the primary driver of our early effects. Nonetheless, we have to interpret the data patterns in the first sentence with caution; they are surprising in light of, and contrary to, prior studies on anticipatory eye movements in language comprehension (e.g., Altmann & Kamide, 2007, 2009). We had interpreted anticipatory looks in these studies as being towards whatever object afforded the action denoted by the verb – e.g. a cake affords eating, and a beer affords drinking (see also Altmann & Kamide, 2007; Chambers, et al., 2004). In the studies reported here, the end-state entailed by the action denoted by the verb (the chopped onion or the broken egg) attracted attention more than did the intact state (again, modulated by that verb – that is, only when the verb implied a change of state). Indeed, eye movements were preferentially directed towards the object that, in general, did *not* afford the action denoted by the verb. Further research is required to establish whether this reflects an over-riding principle, such as looks towards the end-state or goal state associated with the action denoted by the verb. If so, looks towards the initial state (which received more looks than the distractors in Experiments 1 and 2 above) could have been due either to action affordances as originally believed (the initial state is the state that is acted upon in order to achieve the goal state), or due to semantic overlap between the mental representation of the end state and the depiction of the initial state. Further, looks to the chopped onion on hearing “chop” might have been due to lexically-driven local coherence effects (cf. Kukona, Cho, Magnuson, & Tabor, 2014) – “chopped” is a cohort competitor for “chop” and as such could have engendered looks to anything that could be described as “chopped”. Further research is required to explore this possibility.

Our primary focus, however, was not on the first sentence of each pair but rather on the second. In this regard, we again found convergence across both studies. While the critical manipulation occurred at the start of the second sentence ("but first / and then"), the influence this had on the eye movements was not until towards the end of this sentence, during the final reference to the target object. To use the "onion" example as representative of the findings across both studies: It was only during this final reference, in the "*but first*" condition, that participants switched their attention away from the situationally inappropriate chopped onion and towards the situationally appropriate intact onion (again, this was the pattern observed in both experiments). In principle, participants could have switched their attention to the appropriate intact state earlier, or at the very least (in the "*but first*" condition), switched their attention away from the situationally inappropriate (chopped) onion. Indeed, previous studies have suggested that real-world knowledge of temporal structure has immediate and lasting impact on sentence processing (Münte, Schiltz, & Kutas, 1998; Nieuwland, 2015). For example, Münte et al (1998) showed that when the temporal sequence of sentences mismatched the real-world (before vs. after), there were more negative brain responses, as early as 300 ms into processing of the first word of the sentence (e.g., "*Before/After, the scientist submitted the paper, the journal changed its policy*"). That participants did not move their attention away from the situationally inappropriate chopped onion is also at odds with Altmann & Mirkovic's (2007) claim that sentence processing is, generally, maximally incremental – the claim that if there is some information that can be deployed during sentence processing to help anticipate what may come next,

it will be deployed to do exactly that.

There are a number of reasons why participants may have maintained their attention on the (inappropriate) chopped onion: First, there was nothing in the language, from the onset of the temporal expression "*and then/but first*" to the onset of the subsequent noun phrase, to “pull” attention away from whatever was being looked at; and second (and related), this temporal context did not constrain which objects might be referred to next, in which case there was no basis for anticipation and/or attention switching – that is, there was no basis to assume that the onion introduced in the first sentence would or would not be referred to again after the connective, beyond the probabilistic contingencies introduced by the trial structure across the experiment. In fact, participants could have deduced that whenever there were two objects of the same kind in the display, one of which was in a different state to the other, *if* one of those objects was referred to in the first sentence, the other would be referred to after a “*but first*”. But participants were apparently not tracking the trial structure so closely.

While there may not have been anything definitive to shift attention towards, there was very definitely a reason to assume that the depiction of the changed object, at "*but first*" was not situationally appropriate in the new situation that was about to be described: Temporal connectives such as "*but first*" or "*but while*" indicate a situational shift. So in light of the actual impossibility of the changed object even existing during this other situation, why did the eyes still remain fixated on that object? There is another possibility to explain why the eyes lingered on this situationally inappropriate

object state: Yes, it was situationally inappropriate, but that does not mean it was discourse-irrelevant – if it was, why mention it at all? In the absence of any definitive referent that could be anticipated after "*but first* …" (or even after "*but first she will weigh*…") the chopped onion is still presumed relevant to the broader discourse (again, if it wasn't, why mention it or the event it took part in?), and as such the chopped onion may have remained the most active representation in the discourse (beyond the woman who did the actual chopping, although she was not actually depicted in the displays). Hence those lingering looks – they may have reflected the presumed discourse-relevance of the (presumed to be) temporarily inappropriate representation.

Taken together, the results from Experiments 1 and 2 demonstrate that attention is (eventually) guided towards the situationally appropriate depiction of object state. We interpret the results as evidence of the mapping between the actual onion or egg previously referred to, the intended referent selected from among the available mental representations, and the featural overlap between this representation and the visual depictions of the possible alternatives. As participants' attention was explicitly drawn in the language, at the end of the second sentence, to the initial state of the onion or egg, participants had in their situation model of the unfolding event a representation of those initial states, and this representation better matched the visual depictions of an intact onion and an intact egg depending on the language context. Conversely, the chopped onion and broken egg were poorer matches against such a mental representation in the "but first" conditions. In other words, eye movements towards the intact object at the end of the second sentence may not have reflected a commitment to that object being a depiction of the

actual onion or egg that had been acted upon. Instead, eye movements most likely reflected the overlap between the mental representations of the target object that was described in language and the actual visual representations depicted in the visual scene.

Conclusion

To conclude, the two experiments presented here have demonstrated that language-mediated eye movements are sensitive to the mapping between mental object-state representations and object state depictions in a concurrent visual scene. The time-course with which visual attention is mediated by this mapping suggests the mental equivalent of putting a representation "*to one side*"; after "*she chopped the onion, but first*…" the onion in its chopped state is no longer situationally relevant, yet if it were not relevant *at all* (that is, its chopping and consequent change of state were not relevant) it would not have been referred to. So pragmatically, it is likely that it will be referred to again. Hence the pragmatic equivalent of "putting it to one side". Further research is required to elucidate the time-course of object updating (i.e. the updating of its state); the conditions that may modulate this time-course; and the interaction between this time-course and visual attention in the visual world paradigm (VWP). Nonetheless, the data demonstrate that language-mediated eye movements can be used in studies intending to explore the dynamics with which object representations are updated, as or after they become outdated. Further studies are required to examine what factors contributed to driving participants' eye movements towards one depicted state representation or another, and whether such eye movements reflect binding of the distinct state depictions to one another (despite that they

are mutually exclusive in space and time) or whether they reflect instead (or as well) semantic overlap between mental object-state representations and the depicted state representations. Regardless of which it is, our data suggest a fruitful methodological and theoretical approach to understanding the dynamics of object updating during sentence comprehension.

Acknowledgements

We thank Dr. Silvia Gennari and Dr. Shirley-Ann Rueschemeyer for their comments when these studies were conducted. We also thank two anonymous reviewers for their insightful comments.

Funding information

This research was supported by an Overseas Graduate Student Research Scholarship from the University of York awarded to XK, and a research grant from the Economic and Social Research Council (ESRC, ref. RES-062-23-2749) awarded to GTMA.

Author statement

Xin Kang: methodology; project administration; design; data collection; data analysis; data curation; writing, including original draft & review & editing, visualization. **Gitte H. Joergensen**: methodology; project administration; data collection; writing, including original draft & review & editing. **Gerry T.M. Altmann:** conceptualization; design; writing, including original draft & review & editing.

Competing interests

The authors declare that they do not have any competing interest.

Data accessibility statement

The materials and data of all experiments can be found on the Open Science Framework Folder (https://osf.io/rgz62/)

References

Altmann, G. T. M., & Ekves, Z. (2019). Events as intersecting object histories: A new theory of event representation. *Psychological Review, 126*(6), 817-840.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition, 73,* 247–264.

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*(4), 502–518.

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

Altmann, G. T. M, & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*. *33*(4), 583-609.

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.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software Articles*, *67*(1), 1–48.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30 (3), 687-696.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review*, *12*(3), 453–459.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*(4), 547–582.

Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. J*ournal of Memory and Language, 26,* 69–83.

Hindy, N. C., Altmann, G.T.M., Kalenik, E., & Thompson-Schill, S.L. (2012). The effect of object-state changes on event processing: Do objects compete with themselves? *The Journal of Neuroscience, 32(17)*, 5795–5803.

Hoover, M. A. & Richardson, D. C. (2008). When Facts Go Down the Rabbit Hole: Contrasting Features and Objecthood as Indexes to Memory. *Cognition 108(2)*, 533-42.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96(1),* B23-B32.

Huettig, F., & Altmann, G. T. M. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition, 15(8)*, 985-1018.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge, MA: Harvard University Press.

Kalénine, S., Mirman, D., Middleton, E. L., & Buxbaum, L. J. (2012). Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(5), 1274–1295.

Kamide, Y., Altmann, G. T. M., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: evidence from anticipatory eye-movements. *Journal of Memory and Language, 49,* 133–156.

Kamide, Y., Lindsay, S., Scheepers, C., & Kukona, A. (2016). Event processing in the visual world: Projected motion paths during spoken sentence comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(5), 804–812.

Kang, X., Eerland, A., Joergensen, G. H., Zwaan, R. A., & Altmann, G. T. M. (2019). The influence of state change on object representations in language comprehension. *Memory & Cognition*. https://doi.org/10.3758/s13421-019-00977-7

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events

on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language*, *57*(4), 519–543.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition, 95,* 95–127.

Kukona, A., Altmann, G. T. M., Kamide, Y. (2014). Knowing what, where, and when: Event comprehension in language processing. *Cognition, 133* (1), 25-31.

Kukona, A., Cho, P. W., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(2), 326–347.

Kuperberg, G. R., & Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? *Language, Cognition, and Neuroscience, 31:* 32-59.

Lenth, R. (2018). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* R package version 1.1.3. https://CRAN.R-project.org/package=emmeans.

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass, 3(6)*, 1417-1429.

Münte, T. F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, *395*(6697), 71–73.

Nieuwland, M. S. (2015). The Truth Before and After: Brain Potentials Reveal Automatic Activation of Event Knowledge during Sentence Comprehension. *Journal of Cognitive Neuroscience*, *27*(11), 2215–2228.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Radvansky, G. A. (2009). Spatial directions and situation model organization. *Memory & Cognition, 37,*796-806.

Radvansky, G. A. (2012). Across the event horizon. *Current Directions in Psychological Science, 21*, 269-272.

Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory & Cognition, 34(5),* 1150-1156.

Radvansky, G. A. & Copeland, D. E. (2010). Reading times and the detection of event shift processing. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 36,* 210-216.

Solomon, S. H., Hindy, N. C., Altmann, G. T. M., & Thompson-Schill, S. L. (2015). Competition between Mutually Exclusive Object States in Event Comprehension. *Journal of Cognitive Neuroscience*, *27*(12), 2324–2338.

Speer, N. K., & Zacks, J. M. (2005). Temporal changes as event boundaries: Processing and memory consequences of narrative time shifts. *Journal of Memory and Language, 53,* 125-140.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12,* 153-156.

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, *138* (2), 236–257.

Yaxley, R. H. & Zwaan, R. A. (2007). Simulating visibility during language. *Cognition, 150,* 229-236.

Yee, E., & Sedivy, J. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*(1), 1-14.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind/brain perspective. *Psychological Bulletin, 133*, 273–293.

Zwaan, R. A. (1996) Processing narrative time shifts. Journal of Experimental

Psychology: *Learning, Memory, and Cognition. 22 (5)*: 1196-1207.

Zwaan, R. A., & Madden, C. J. (2004). Updating situation models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 283–288.

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

Zwaan, R. A. & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123 (2)*: 162-185.

Zwaan, R. A., Stanfield, R. A., Yaxley, R. H. (2002). Do language comprehenders routinely represent the shapes of objects? *Psychological Science, 13,* 168-171.