ELSEVIER

Cognition 73 (1999) 247–264

COGNITION

www.elsevier.com/locate/cognit

Incremental interpretation at verbs: restricting the domain of subsequent reference

Gerry T.M. Altmann*, Yuki Kamide

Department of Psychology, University of York, Heslington, York YO10 5DD, UK

Received 12 March 1999; received in revised form 13 August 1999; accepted 10 September 1999

Abstract

Participants' eye movements were recorded as they inspected a semi-realistic visual scene showing a boy, a cake, and various distractor objects. Whilst viewing this scene, they heard sentences such as '*the boy will move the cake*' or '*the boy will eat the cake*'. The cake was the only edible object portrayed in the scene. In each of two experiments, the onset of saccadic eye movements to the target object (the cake) was significantly later in the *move* condition than in the *eat* condition; saccades to the target were launched after the onset of the spoken word *cake* in the *move* condition, but before its onset in the *eat* condition. The results suggest that information at the verb can be used to restrict the domain within the context to which subsequent reference will be made by the (as yet unencountered) post-verbal grammatical object. The data support a hypothesis in which sentence processing is driven by the predictive relationships between verbs, their syntactic arguments, and the real-world contexts in which they occur.

Keywords: Parsing; Thematic roles; Eye movements

1. Introduction

Most theories of language comprehension assume, at least tacitly, a distinction between the identification of the entities taking part in the event (or state) described by a fragment of the language, the identification of the roles that are played out in that event, and the assignment of specific roles to specific entities. For example, on hearing a sentence such as '*the boy will pick up the ornate red vase*', the meanings of the noun phrases *the boy* and *the ornate red vase* determine which things in the

* Corresponding author. Tel.: +44-1904-434-362; fax: +44-1904-433-181.
E-mail address: g.altmann@psych.york.ac.uk (G.T.M. Altmann)

0010-0277/99/$ - see front matter
PII: S0010-0277(99)00059-1

(mental) world are taking part in the event under consideration. The meaning of the verb *pick up* defines the relationship between the thing that will do the picking up (the agent), and the thing that will be picked up (the theme); and knowledge of the grammar of the language determines which positions in the sentence are associated with which roles (if the language uses positional information to convey such information), and consequently which entities referred to in the language fill which specific roles. Many theories also now assume that knowledge of the roles associated with the action denoted by a verb is represented together with knowledge of where within the sentence to locate the expressions that will receive those roles. That is, it is assumed that aspects of grammatical knowledge are lexicalised and accessed together with other aspects of a verb's meaning (Bresnan, 1982; MacDonald, Pearlmetter & Seidenberg, 1994).

In this paper, we demonstrate that information extracted at verbs not only serves to identify roles and the positions within the sentence where the recipients of those roles can be identified, but can on occasion serve to identify directly the (real or mental world) entities that play out those roles – we demonstrate that information extracted at a verb can function in much the same way as the information extracted at, for example, adjectives like *ornate* or *red*, or nouns such as *vase*.

It is now well-established that the processing of referring expressions such as '*the ornate red vase*' proceeds incrementally – the adjectives *ornate* and *red* provide constraints on the range of entities denoted by the subsequent noun *vase*, and as each adjective is encountered, so the constraints it conveys are used to further refine the set of referents which satisfy the accumulating constraints (Altmann & Steedman, 1988; Sedivy, Tanenhaus, Chambers & Carlson, 1999). On hearing an expression such as '*pick up the ornate red vase*', participants looking at a visual scene containing an ornate red vase will initiate eye movements to the vase as soon as the accumulating constraints identify a unique referent (Eberhard, Spivey-Knowlton, Sedivy & Tanenhaus, 1995; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995; Sedivy, Tanenhaus, Chambers & Carlsin, 1999). Thus, reference to (visual world) context can be achieved in an incremental, piecemeal manner, using information to narrow down the set of available referents as soon as that information first becomes available – even before the head noun (*vase*, in the example above) is itself encountered.

One issue with regard to the application of accumulating constraints concerns the provenance of these constraints. Although referring expressions do generally convey sufficient information with which to uniquely identify the intended referent, there are occasions when verbs convey such uniquely identifying information also. Altmann (1999) describes a study (Experiment 1) in which participants read sentences beginning, for example, '*the boy ate the …*' and had to judge, as each word appeared across the screen, whether the sentence did or did not make sense up to that word. These sentences were preceded by a context sentence which either introduced something edible into the story, or did not. When it did not, there were significantly more 'no' judgements at the verb in the following target sentence than when it did. It was argued, from these data, that the processor can project, at the verb, the upcoming referring expression in direct object position, and that on doing

so, it evaluates that projected expression with respect to the context and the entities within that context which fulfil the selectional restrictions of the verb (with 'no' judgements arising when there were no entities that satisfied those restrictions). On this view, if there is just one edible thing in the context, and the processor encounters a sentence fragment such as '*the boy will eat*', it should act little differently from the situation where it encounters the fragment '*pick up the ornate*' – in the *ornate* case, the adjective narrows down the set of available referents according to which ones satisfy the constraint of being ornate. In behavioural terms, when the fragment is heard in the context of a visual scene containing just one ornate thing, the application of this constraint results in eye movements to that one ornate thing within just a few hundred milliseconds of the onset of the word *ornate*. In the *eat* case, if a sentence fragment such as '*the boy will eat*' were to be heard in the context of a visual scene containing just one edible thing, the application of the must-be-edible constraint should result in eye movements to that one edible thing soon after verb onset. The experiments reported below test precisely this prediction: that semantic information extracted at the verb is able to guide visual attention towards an appropriate object in visual context (as determined by that semantic information) even before the semantic properties of the direct object become available. Unlike the word-by-word stop-making-sense judgement task employed by Altmann (1999), the experiments we report below did not require any artificial segmentation of the linguistic input, and in the case of Experiment 2, did not require participants to do anything other than look-and-listen.

The methodology we adopted for this study, and which was used also by Eberhard et al. (1995) and Sedivy et al. (1999), is based on an early observation by Cooper (1974), who pointed out that when participants are simultaneously presented with spoken language whilst viewing a visual scene, their eye movements are very closely synchronised to a range of different linguistic events in the speech stream. The linguistic sensitivity of this technique has been validated recently in studies by Allopenna, Magnuson & Tanenhaus (1998), Eberhard et al. (1995) and Sedivy et al. (1999). For example, in addition to the demonstration that eye movements are closely synchronised to the referential processing of the concurrent linguistic input (Eberhard et al., 1995; Sedivy et al., 1999; see above), Allopenna et al. (1998) demonstrated that the probability of fixating one object amongst several when hearing that object's name is closely related, and the fixations closely time-locked, to phenomena associated with auditory word recognition. This temporal sensitivity to both word identification and subsequent referential processing makes the methodology ideally suited to our investigation of verb-mediated referential processing.

2. Experiment 1

2.1. Participants

Twenty-four participants from the University of York student community took

part in this study. They participated either for course credit or for £2.00. All were native speakers of English and either had uncorrected vision or wore soft contact lenses or spectacles.

2.2. Stimuli

Sixteen sets of stimuli were devised each consisting of a single semi-realistic visual scene and two accompanying sentences (see Appendix A and Fig. 1). The visual scenes were created using commercially available ClipArt packages. The scenes were constructed using a 16-colour palette, and were presented on a 17″ viewing monitor at a resolution of 640 × 480 pixels. To describe one scene in detail: it showed a young boy sitting on a floor around which were various items. These were a toy train set, a toy car, a balloon, and a birthday cake. For this scene, two sentences were recorded: '*the boy will move the cake*' and '*the boy will eat the cake*'. For each visual scene, one of the corresponding sentences contained a verb whose selectional restrictions dictated that only a single object in the visual scene could be referred to post-verbally, and the other sentence contained a verb which permitted at least four of the visual objects, including the target object, to be referred to post-verbally. In each case, there was one target object in the visual scene (the cake, in this example), and either three distractor objects (for half the scenes) or four (for the other half). Neither the target object nor the referent of the sentential subject (the boy, in this example) were counted as distractors. In the case of the scenes containing four distractors, one of these could only implausibly be referred to post-verbally,

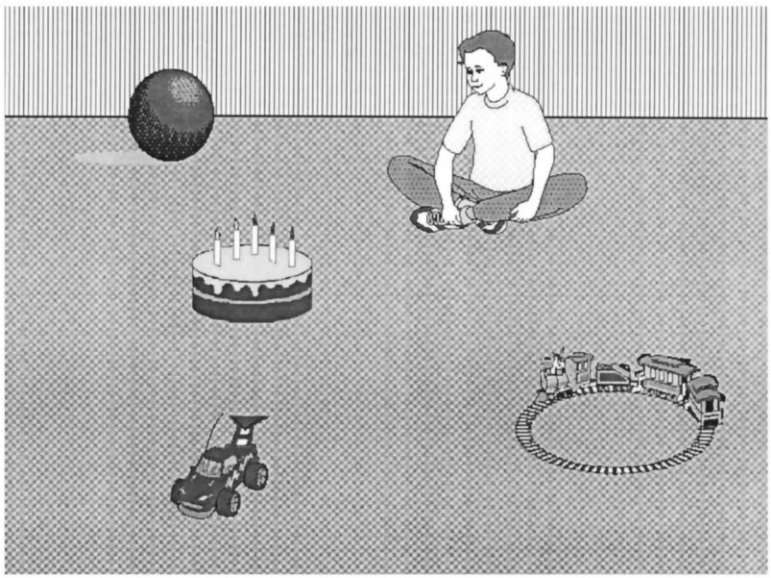

Fig. 1. Example scene used in Experiments 1 and 2 (Sections 2 and 3). Participants heard '*The boy will move the cake'* or '*The boy will eat the cake*' whilst viewing this scene.

irrespective of the verb used. The materials and a description of the objects in each scene are given in Appendix A.

A further 16 sets of stimuli were devised which served as filler items. The post-verbal direct objects in the spoken sentences did not have any corresponding referents in the accompanying visual scenes. There were four versions each of four types of filler: the action denoted by the verb could apply to none of the visual objects; to just one of them; to two of them; or to three of them. In each case, the subsequent direct object would not refer to any of the items contained within the scene. The fillers were designed in this way to enable us to employ the equivalent of a picture verification task in which participants would respond either 'yes' or 'no' (see below, Section 2.3). Two lists of stimuli were constructed, with each participant seeing each scene but hearing only one of the two possible versions of the spoken description that could accompany that scene.

2.3. Procedure

Participants were seated in front of a 17″ display and wore an SMI EyeLink head-mounted eye-tracker, sampling at 250 Hz from the right eye (viewing was binocular). Participants were seated with their eyes between 20″ and 25″ from the display. Their head movements were unrestricted. Participants were instructed to judge whether the sentence they heard could in principle apply to the picture. They were given the example '*the person will light the fire*' and were told to respond 'yes' if the picture showed a fireplace, and 'no' if it did not. No mention was made of the speed with which they should respond. There were two practice trails before the main experimental block. Between each trial, participants were shown a single centrally-located dot on the screen which they were asked to fixate prior to a fixation cross appearing in this position (this procedure allowed recalibration of the eye-tracker). Participants would then press a response button for the next presentation. The onset of the visual stimulus coincided with the onset of the spoken stimulus (both were stored and played from disk). When participants responded ('yes' or 'no' on the button box), the visual stimulus was terminated (in no case did participants respond prior to the end of the spoken stimulus). After every fourth trial, and before the practice and experimental blocks, the eye-tracker was recalibrated using a nine-point fixation stimulus. The EyeLink software automatically validates calibrations and the experimenter could, if required, repeat the calibration process if validation was poor. Calibration took approximately 20 s, and the entire experiment lasted approximately 20 m.

2.4. Results

First, we describe the procedure for analysing the eye-movement data generated by the EyeLink system. *X*–*Y* coordinates output by EyeLink were converted to codes for whichever object lay at those coordinates (fixations beyond 3 or 4 pixels from the object's outermost contour were not deemed as fixations on that object). The background was coded as a separate object. Our primary interest was in determining when, relative to verb onset, the participant's eyes first moved to the target object.

Table 1
Experiment 1 (Section 2): onset of first saccade to the target object in both the 'eat' and 'move' conditions, relative to verb onset, verb offset, determiner onset and noun onset (timings in ms)

	Eat	Move	Difference (move − eat)
Verb onset	611	838	227
Verb offset	228	415	187
Determiner onset	37	234	197
Noun onset	−85	127	212

Markers had been placed in each speech file at verb onset, verb offset, post-verbal determiner onset, post-verbal noun onset, and post-verbal noun offset. This information was entered into the EyeLink output file in real time during each stimulus presentation. We thus had a full record of eye movements relative to these points in the speech wave.

We eliminated from the analysis any saccadic movement to the target object whose onset was prior to the verb's onset – if a participant had been fixating on the target object (the cake) at verb onset, we took the onset of the *next* saccadic movement for the purposes of calculating the onset of the 'first saccade' to the target object. This occurred on approximately 10% of trials. Our rationale was simply that any saccadic movement initiated before verb onset could not possibly have been mediated by information extracted at the verb.

We calculated for each verb (*eat* or *move*), and for each 50-ms interval from the onset of the verb, the cumulative probability across trials of fixating either the target object (the cake), or one of the distractor objects (we calculated the probabilities for each distractor object separately, and then averaged these)[1]; these data are plotted in Fig. 2. We also calculated the onset of the first saccade to the target object relative to both the onset and offset of the verb, to the onset of the post-verbal determiner, and to the onset of the target noun (*cake*). These data are summarised in Table 1. Table 2 summarises the mean duration of the verbs across the two conditions, the mean duration of the intervening determiner (and also the intonational break between verb and determiner), and the mean delay between onset of the verb and onset of the target noun. None of these differed significantly across the two conditions (duration of verb: $F(1,15) = 2.7$, MSE $= 14\,238$, $P > 0.1$; post-verbal intonational break: $F(1,15) < 1$; determiner: $F(1,15) = 1.8$, MSE $= 1625$, $P > 0.2$; verb and break and determiner combined: $F(1,15) < 1$).

Participants fixated the target object post verb-onset on 90% of trials (92% in the *move* condition, and 88% in the *eat* condition – $F1(1,23) = 1.5$, MSE $= 271$, $P > 0.2$; $F2(1,15) = 3.5$, MSE $= 176$, $P > 0.1$). The first saccade to the target object in the *move* condition was launched before noun onset on 38% of all trials,

[1] The data reported in Experiments 1 and 2 (Sections 2 and 3) include the fixation data for distractors that were implausible in either condition. Analyses with these data removed yielded patterns that were statistically the same as those reported here; we report the inclusive data because these more accurately reflect the overall patterns of eye movements observed in the experiments.

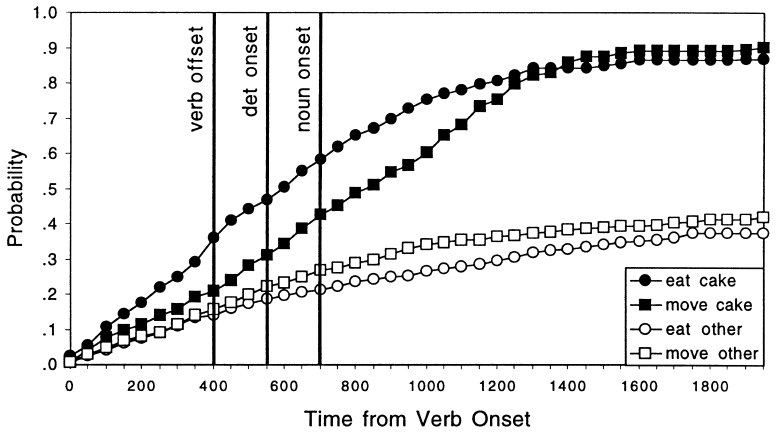

Fig. 2. The cumulative probability of fixating the target object (cake) or a distractor object (other) as a function of condition ('eat' vs. 'move') in Experiment 1 (Section 2). Note: The verb offset, determiner onset, and noun onset are shown, for display purposes, averaged across trials, and are aligned to the 50ms bin within which they fall.

compared to 54% for the *eat* condition. First saccades to a distractor object occurred before noun onset on 26% of trials in the *move* condition, and 20% in the *eat* condition. A two-way ANOVA on the arcsine transformed probabilities (the same patterns were found on the untransformed data) revealed that there were significantly more pre-noun first-looks to the target object than to any of the distractors ($F1(1,23) = 81.8$, MSE $= 1.43$, $P = 0.0001$; $F2(1,15) = 9.9$, MSE $= 0.91$, $P < 0.007$). There was no effect of verb type (*eat* vs. *move*; $F1(1,23) = 2.8$, MSE $= 0.07$, $P > 0.1$; $F2(1,15) = 2.8$, MSE $= 0.07$, $P > 0.1$), but there was an interaction between object type (target or distractor) and verb type ($F1(1,23) = 8.7$, MSE $= 0.33$, $P < 0.008$; $F2(1,15) = 8.2$, MSE $= 0.28$, $P < 0.02$). Planned comparisons revealed that this interaction was due to an effect of verb type on the incidence of first-looks to the target object ($F1(1,23) = 9.7$, MSE $= 0.36$, $P < 0.005$; $F2(1,15) = 9.0$, MSE $= 0.32$, $P < 0.009$) in the absence of an effect of verb type on the incidence of first-looks to the distractors ($F1(1,23) = 1.1$, MSE $= 0.04$, $P = 0.3$; $F2(1,15) = 1.1$, MSE $= 0.04$, $P > 0.3$). Analyses on first-looks prior to verb offset revealed a similar pattern, with marginally more first-looks to the target ($F1(1,23) = 19.1$, MSE $= 0.43$, $P = 0.0002$; $F2(1,15) = 2.0$, MSE $= 0.20$, $P > 0.1$), no effect of verb type ($F1(1,23) = 1.9$, MSE $= 0.05$, $P > 0.1$; $F2(1,15) = 1.1$, MSE $= 0.02$, $P > 0.3$), and an interaction between verb and object type that approached significance ($F1(1,23) = 3.3$, MSE $= 0.10$, $P < 0.09$; $F2(1,15) = 4.2$, MSE $= 0.07$, $P < 0.06$). Planned comparisons revealed an effect of verb type on first-looks to the target object ($F1(1,23) = 4.9$, MSE $= 0.15$, $P < 0.04$; $F2(1,15) = 4.8$, MSE $= 0.07$, $P < 0.05$), but not on first-looks to the distractors ($F1(1,23) < 1$; $F2(1,15) < 1$).

The onset of the first post-verb-onset saccade to the target object in the *move* condition occurred 127 ms after the onset of the target noun. In the *eat* condition the

Table 2
Word durations for the 'eat' and 'move' sentences (timings in ms)

Duration	Eat	Move	Difference (move − eat)
Verb	383	423	40
Post-verbal break	192	180	−12
Determiner	122	107	−15
Verb + break + determiner	697	710	13

onset occurred 85 ms *before* the onset of the target noun – $F1(1,23) = 12.3$, $\text{MSE} = 541\,840$, $P < 0.002$; $F2(1,15) = 11.9$, $\text{MSE} = 342\,658$, $P < 0.004$. An ANCOVA with verb duration as a covariate revealed a significant effect of verb type ($F2(1,14) = 8.75$, $\text{MSE} = 269\,190$, $P = 0.01$), and no effect of the covariate ($t = 0.3$, $P > 0.7$).

2.5. Discussion

When the verb's selectional restrictions could apply to only one of the objects in the visual scene, the probability of looking at the appropriate object before the onset of the post-verbal noun was significantly elevated compared with the case where the verb's selectional restrictions could apply to more than one object (0.54 and 0.38, respectively). Indeed, this same pattern was observed even earlier, at the offset of the verb (0.29 and 0.22, respectively) (see Fig. 2). The figure also illustrates the finding that there were more first-looks, prior to noun-onset, to the target object than to any other distractor object, even in the *move* condition. We presume that this reflects the fact that not all the distractors were equally plausible as objects for the 'non-selecting' verb (they were not equally plausible in respect of being moveable, for example – see Appendix A for examples)[2], or alternatively, that incidental properties of the target objects may have influenced their visual salience relative to the distractors (including colour, shape, proximity to the agent of the action denoted by the verb, and so on). Nonetheless, the significant effect of verb type on the probability of first-fixating the target object indicates mediation of the eye-movements by linguistic factors.

The mean delay between the verb's onset and the launch of the first saccadic movement to the target object was considerably shorter when the verb's selectional restrictions 'picked out' that object than when they did not – in the former case, the first saccade was launched well before the onset of the post-verbal noun, and in the

[2] In collaboration with Sarah Haywood, we presented 50 participants with each of the scenes used in Experiments 1 and 2 (Sections 2 and 3) and told them that we wished to know what was most likely to happen next. We asked them, therefore, to complete a short sentence fragment such as '*the boy will move the*'. The fragments were created from the sentences used in the *move* conditions of Experiments 1 and 2 (Sections 2 and 3). The completions indicated, overall, that the target object (e.g. the cake) was twice as likely to be referred to in direct object position as was any one of the distractors (e.g. the ball, the toy train, or the toy car) – indicating that even though these distractors were indeed moveable (or equivalent), participants judged the target object as more plausibly moved.

latter case, it was launched well after that onset (−85 and 127 ms, respectively). The (non-significant) 40 ms difference in the duration of the verbs did not account for this 227 ms difference in launch time relative to the verb onset. According to some estimates (e.g. Matin, Shao & Boff, 1993), it takes up to 200 ms to program a saccadic eye movement, in which case these data illustrate very fine time-locking between the extraction of verb information from the auditory stream and the launching of eye movements driven by that information (228 ms after verb offset). These data all lead to the same conclusion: information extracted at the verb can be used to guide eye movements to whichever object in the visual context satisfies the selectional restrictions of the verb.

The absolute latencies reported here between verb onset and the onset of the saccadic eye movement to the target object are comparable with figures reported by Eberhard et al. (1995) and by Sedivy et al. (1999). In those studies, participants heard sentences such as '*Touch the plain red square*' or '*Is there a tall glass?*'. In both cases, eye movements were often initiated to the target object before the onset of the head noun when the information conveyed by the prior adjective was sufficient to restrict the domain of reference to just one object in the visual scene – eye movements were initiated within around 550 ms of the onset of *plain* when there were several different objects of which just one was plain (Eberhard et al., 1995), and within around 650 ms of the onset of *tall* when there was one prototypically tall thing (as well as a prototypically short thing – Sedivy et al., 1999). In both cases, longer latencies were recorded when the adjective was not effective at restricting the domain of reference to a single object. In the study described above, eye movements were initiated, in the *eat* condition, 611 ms after verb onset. Our data suggest therefore that information extracted at the verb can have behavioural consequences that are virtually identical to the situations described by Eberhard et al. (1995) and Sedivy et al. (1999) – that is, information extracted at the verb can drive eye movements to a particular object in visual context in much the same way as can information extracted at a prenominal adjective.

It is conceivable, however, that our data reflect the exigencies of the judgement task – normal language comprehension does not usually require meta-linguistic judgements, and the requirement to make such judgements may have induced anticipatory processing strategies which do not reflect normal processing. To address this issue, we repeated Experiment 1, but without explicitly asking participants to perform any meta-linguistic judgement.

3. Experiment 2

3.1. Participants

Twenty-four participants from the University of York student community took part in this study. They participated either for course credit or for £2.00. All were native speakers of English, had either uncorrected vision or wore soft contact lenses or spectacles, and had not taken part in the previous experiment.

3.2. Stimuli

The same stimuli were used as had been used in Experiment 1 (Section 2.2).

3.3. Procedure

The procedure was identical to that used in Experiment 1 (see Section 2.3) with one difference: participants were informed that each picture would be accompanied by a short sentence, but that 'Each picture will be accompanied by a short sentence spoken over the loudspeakers, but in this version of the experiment we aren't asking you to pay any particular attention to the sentences (some refer to the things in the pictures, others don't, but that isn't relevant to this experiment) '.

3.4. Results

Participants fixated the target object on 93% of trials (93% in the *move* condition, and 93% in the *eat* condition – both $F < 1.0$). The first saccade to the target object in the *move* condition was launched before noun onset on 18% of all trials. The equivalent saccade in the *eat* condition was launched before noun onset on 32% of trials. The figures for the distractor objects were 12 and 15%, respectively (see Fig. 3). A two-way ANOVA on the arcsine transformed probabilities revealed that there were marginally more pre-noun first-looks to the target object than to any of the distractors ($F1(1,23) = 9.7$, MSE $= 0.21$, $P < 0.005$; $F2(1,15) = 1.9$, MSE $= 0.12$, $P > 0.1$). There was a significant effect of verb type ($F1(1,23) = 11.4$, MSE $= 0.46$, $P < 0.003$; $F2(1,15) = 20.2$, MSE $= 0.35$, $P = 0.0004$), and a marginally significant interaction between object type (target or distractor) and verb type ($F1(1,23) = 2.1$, MSE $= 0.10$, $P > 0.1$;

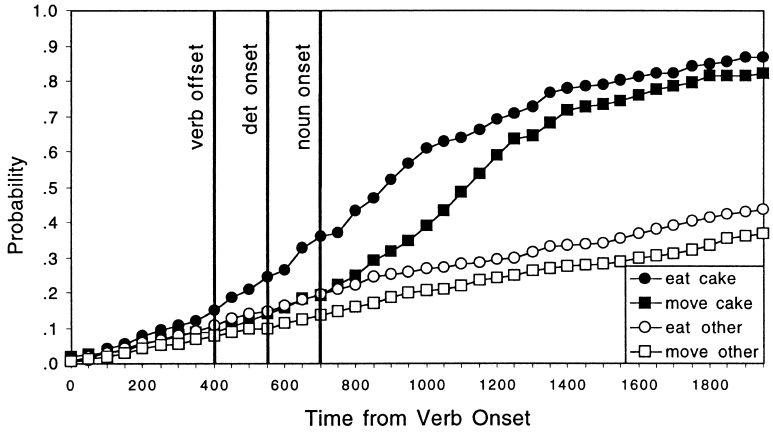

Fig. 3. The cumulative probability of fixating the target object (cake) or a distractor object (other) as a function of condition ('eat' vs. 'move') in Experiment 2 (Section 3). Note: The verb offset, determiner onset, and noun onset are shown, for display purposes, averaged across trials, and are aligned to the 50ms bin within which they fall.

$F2(1, 15) = 16.1$, MSE $= 0.13$, $P < 0.002$). Planned comparisons revealed an effect of verb type on the incidence of first-looks to the target object ($F1(1, 23) = 10.9$, MSE $= 0.49$, $P = 0.003$; $F2(1, 15) = 57.1$, MSE $= 0.44$, $P = 0.0001$) in the absence of an effect of verb type on the incidence of first-looks to the distractors ($F1(1, 23) = 1.5$, MSE $= 0.07$, $P > 0.2$; $F2(1, 15) = 3.5$, MSE $= 0.03$, $P > 0.08$). Analyses on first-looks prior to determiner onset revealed more first-looks following *eat* than following *move* ($F1(1, 23) = 9.5$, MSE $= 0.23$, $P < 0.006$; $F2(1, 15) = 9.5$, MSE $= 0.17$, $P < 0.008$), but no effect of object type ($F1(1, 23) < 1$; $F2(1, 15) < 1$), and a marginally significant interaction between verb and object type ($F1(1, 23) < 1$; $F2(1, 15) = 5.9$, MSE $= 0.07$, $P < 0.03$). Planned comparisons revealed an effect of verb type on first-looks to the target object ($F1(1, 23) = 5.1$, MSE $= 0.20$, $P < 0.04$; $F2(1, 15) = 20.6$, MSE $= 0.23$, $P = 0.0004$), but not on first-looks to the distractors ($F1(1, 23) = 1.3$, MSE $= 0.05$, $P > 0.2$; $F2(1, 15) = 1.2$, MSE $= 0.01$, $P > 0.2$).

The onset of the first post-verb-onset saccade to the target object in the *move* condition occurred 536 ms after the onset of the target noun. In the *eat* condition the onset occurred 291 ms after the onset of the target noun – $F1(1, 23) = 8.1$, MSE $= 721\,280$, $P < 0.01$; $F2(1, 15) = 15.6$, MSE $= 492\,452$, $P < 0.002$ (see Table 3). This difference was also significant in an ANCOVA ($F2(1, 14) = 13.57$, MSE $= 453\,989$, $P = 0.002$) which revealed that it was not due to any differences in verb durations across the two verb types ($t = 0.4$, $P > 0.7$).

3.5. Discussion

In Experiment 2 (Section 3), like in Experiment 1 (Section 2), the probability of first-fixating the target object between verb onset and noun onset was greater in the *eat* condition than in the *move* condition. Fig. 3 shows that this difference begins to manifest itself just after verb offset, and indeed, by the time the onset of the determiner was encountered, the difference was significant. Allowing time to program the saccadic movement means that this difference, although manifest just after verb offset, is most likely due to differences in processing that occurred during the acoustic lifetime of the verb. However, whereas in this experiment, the difference begins to manifest itself just after verb offset, in Experiment 1 (Section 2), the difference manifested itself somewhat sooner; during the lifetime of the verb itself. Overall, it would appear that the same pattern of anticipatory eye movements was observed in Experiment 2 (Section 3) as in Experiment 1 (Section 2); in Experiment

Table 3
Experiment 2 (Section 3): onset of first saccade to the target object in both the 'eat' and 'move' conditions, relative to verb onset, verb offset, and noun onset (timings in ms)

	Eat	Move	Difference (move − eat)
Verb onset	988	1246	258
Verb offset	605	823	218
Determiner onset	413	643	230
Noun onset	291	536	245

2 (Section 3), however, the patterns appears to have shifted 'downstream' by around 350 ms. Nonetheless, and despite some evidence of a strategic influence on the timing of linguistically-mediated saccadic eye movements, the data provide clear evidence, in the absence of an explicit meta-linguistic task, for the same qualitative effect as seen in Experiment 1 (Section 2): information extracted at the verb can guide eye movements to whichever object in the visual context satisfies the selectional restrictions of the verb, and these movements can be initiated prior to the onset of the spoken word referring to that object.

4. General discussion

Our data demonstrate that information extracted at the verb can be used to guide eye movements to whichever object in the visual context satisfies the selectional requirements of the verb. This guidance is initiated before the linguistic expression corresponding to the verb's direct object is encountered. And although we have some evidence for faster guidance as a function of the exigencies of the task, the same pattern of early saccadic movements (with launch of the saccadic movement taking place prior to the onset of the critical referring expression) is seen even when participants are asked, in effect, to ignore the auditory stimulus. Nonetheless, the significance of an equivalent pattern irrespective of whether participants were required to make an explicit judgement should be judged with some degree of caution. The identical stimuli were used in both experiments, meaning that in Experiment 2 (Section 3), half the sentential stimuli referred to objects that were not in fact contained within the visual scene (and were designed to elicit 'no' responses in Experiment 1 (Section 2)). It is conceivable that the presence of such sentences prompted participants to develop a strategy over the course of the experiment which was still somewhat artificial – they may have attempted to anticipate whether the sentence would or would not apply to the visual scene, and hence the similar results to Experiment 1 (Section 2). However, if participants developed their own strategy over the course of the experiment, some difference should be observable in the data between earlier and later trials. Comparing patterns in the first and second half of the experiment yielded no significant differences (there were no main effects nor interactions with first versus second) – in fact, the proportions of pre-noun-onset first looks to the target in the *eat* and *move* conditions were identical in the two halves of the experiment (and separate planned comparisons confirmed that the differences in proportions of first looks to the target between the *eat* and *move* conditions were statistically significant; first half: $F1(1,23) = 8.5$, MSE $= 0.19$, $P < 0.008$; $F2(1,7) = 21.6$, MSE $= 0.20$, $P < 0.003$; second half: $F1(1,23) = 10.5$, MSE $= 0.23$, $P < 0.004$; $F2(1,7) = 39.1$, MSE $= 0.25$, $P = 0.0004$). Nonetheless, we cannot rule out the possibility that (some) participants may have interpreted the task as requiring some form of (implicit) meta-linguistic judgement which may have contributed to the overall pattern of results we observed.

We do not view this last possibility as a limitation on the generalizability of our data. Sedivy et al. (1999) also employed a judgement task ('*Is there a tall glass?*')

with trials designed to elicit 'no' responses. And in both the Sedivy et al. (1999) and Eberhard et al. (1995) studies, manipulation tasks ('*Pick up the tall glass*') revealed substantially similar results in the absence of any such 'no' trials. It is certainly true that the situation participants found themselves in in Experiment 2 (Section 3) was artificial insofar as they would hear sentences which either did or did not apply to the scenes they were viewing – but it is a situation that is reminiscent, in fact, of natural language processing in everyday contexts.

Our result has important implications for models of sentence processing – it supports the contention that the processor can project, at the verb, an upcoming referring expression in grammatical object position, that it can immediately attempt to establish anaphoric dependencies between that projected expression and the context, and that it can attempt to do so on the basis of the thematic fit between the entities in the context and the verb (Altmann, 1999). Central to this claim is that what is projected is not structure *per se*, but *interpreted* structure.

One issue which is not resolved by these data concerns the consequence of verb-mediated reference. It is possible that information at the verb does nothing more than restrict the domain of (subsequent) reference. Thus, although the programming of eye movements in our study towards the appropriate visual referent was initiated within perhaps as little as 30 ms of verb offset (in Experiment 1 (Section 2)), this need not mean that the processor had already assigned to that visual entity the role associated with whatever was about to be referred to in grammatical object position (in this case, the patient role). However, to direct visual attention to an appropriate referent on the basis of a verb's selectional restrictions does necessitate some evaluation of the thematic fit between the verb and whichever entities are available in the context. Whatever mental representations underlie this evaluation, they must encode the relationship between the verb's meaning and the properties of the things that are available in the context to take part in the action or event denoted by the verb. And if this evaluative process, and the encodings it entails, do not constitute role assignment *per se*, it is unclear to us what more would be encoded if such role assignments (and whatever encodings *they* entail) subsequently took place.

The information that we believe was extracted at the verbs in our studies, and which guided eye movements to the thematically appropriate visual world object, appears to be of the same kind that McRae, Ferretti and Amyote (1997) referred to during their discussion of verb-specific knowledge and thematic role assignment. They view thematic roles as verb-specific concepts that are formed through experience of the entities that play a role in the event to which a verb refers. Thus, thematic roles reflect world knowledge that changes dynamically during language learning as experience is amassed (Altmann, 1997, 1999). McRae et al, (1997) argue that when the verb is encountered, verb-specific knowledge about typical agents and patients (and whatever other roles tend to be associated with the verb) is activated and compared against candidate noun fillers. Thus, role concepts activated at the verb are compared against one or more lexical concepts. The data we have reported here are entirely compatible with the view espoused by McRae et al. (1997), and although our studies were not designed to test directly the details of the proposal of McRae et al., they do take the theoretical claims further. Our data demonstrate that verb-

specific knowledge, once activated, can be compared against candidates for roles that have not yet been syntactically realized (that is, the grammatical positions normally associated with these roles have not yet been encountered), and that these candidates are not linguistic entities but entities existing in the discourse or real-world context – the verb's argument structure permits the unrealized grammatical object to be projected, but because that object is unrealized, no single lexical concept associated with that projected expression can be activated. All that is available at the point of projection is a thematic specification (the role concept, in the terms of McRae et al.) – an abstraction, in effect, of the range of lexical concepts that could subsequently be activated. We thus believe that our data show that there are circumstances when thematic fit is not computed against individual lexical concepts *per se*, but against discourse or real-world entities (as mediated by the mental representation of those entities).

Our belief that thematic fit can be computed against discourse entities as well as real-world entities stems from data reported in Altmann (1999). There, participants read (in Experiment 2) passages such as '*Andrea wrote a card for the infant/politician. She sent a little rattle to him/his baby when she was in Ohio'*. Participants were asked to judge, for each word they read (in a self-paced word-by-word moving window presentation), whether the target sentence continued to make sense. On 30% of trials, *rattle* was judged implausible in the *politician* case compared to just 5% of trials in the *infant* case. The implausibility of *rattle* could only arise, it was argued, if the politician was assumed, by the time '*a little rattle*' was encountered, to be the recipient of whatever was being sent. It was argued that these data reflected the operations of a processor that predictively activates representations at the verb which are evaluated with respect to the thematic fit between the verb and, in this particular case, pre-existing discourse entities. It was further argued that this predictive evaluation process corresponded to role assignment. Thus, by the time '*him/his baby*' was encountered in the example above, the processor would already have assigned the role normally associated with this grammatical position to a pre-existing discourse entity (in which case, presumably, the processing of the referring expression in that position would either confirm or disconfirm a prior assignment, rather than trigger the initial assignment). The present data confirm that representations can be activated at a verb that 'refer' to entities for which a referring expression would normally occur post-verbally, and establish that these representations can guide visual attention towards entities in the (mental representation of the) context which fulfil aspects of the verb's thematic specification (in the experiments reported here, its selectional restrictions). Taken together, the data suggest that the activation of verb-specific knowledge (or concepts) not only serves as the basis for driving attention towards the mental entities that best match those concepts, but also serves as a basis for modulating the way that subsequent linguistic expressions are processed, as in the case of, for example, the Altmann (1999) study.

According to the account of verb-specific knowledge espoused by McRae et al. (1997), the information we have referred to here under the heading of 'selectional restriction' is little different from information concerning the real-world plausibility (or otherwise) of different real-world entities taking part in the event denoted by the

verb – knowledge of both is accumulated through experience of such events and the entities taking part in them. Could real-world plausibility form the basis, then, for verb-mediated reference? For instance, in a visual context showing a woman, a plate of vegetables, and a plate of cat food, the sentence fragment '*she will eat*' may direct the eyes to the vegetables even though the cat food also satisfies the selectional restrictions of the verb *eat*. In this case, the implausibility of the cat food as the object of the subsequent eating may also serve to mediate visual attention. Similarly, in a visual context showing a cat, a plate of vegetables, and a plate of cat food, the fragment '*she will eat*' may direct the eyes now to the cat food. It is an empirical issue whether real-world plausibility can be used in the same way as selectional information – selectional restrictions tend to reflect dependencies between verbs and their objects irrespective of their subjects, whereas plausibility effects reflect contingencies between verbs, their objects, *and* their subjects (so selectional restrictions can be lexicalised in a way that plausibility can not).

There do exist data which speak to this latter issue, albeit indirectly. Chambers, Tanenhaus, Eberhard, Carlson & Filip (1998) investigated the processing of prepositional phrases such as '*inside the can*' in the context of instructions such as '*Pick up the cube. Now put it inside the can*'. In a visual scene containing just one can, Chambers et al. (1998) reported earlier saccadic movements to the can when the preposition was *inside* than when it was *below* (which, given the nature of the visual arrays, was treated as '*in front of*'), and there was some evidence that eye movements to the can were initiated during the preposition itself. In some respects, this case is equivalent to our own manipulation, using prepositions which either do (*inside*) or do not (*below*) select amongst the different entities, instead of, as in our case, verbs which either do (*eat*) or do not (*move*) select amongst different entities. Chambers et al. (1998) also report a case including a small can, a large can, and a bowl, and either a small cube that could be placed inside any of these three containers, or a large cube which could be placed inside only the large can or the bowl. Participants were again asked to pick up the cube and to either '*put it inside the can*' or '*put it inside a can*'. Chambers et al. found earlier saccadic movements to the target container (whichever can participants chose to put the cube inside) when the moved object was the large cube than when it was the small cube. No such difference was found with the indefinite ('*a can*') instruction. At first glance, these data suggest that the preposition *inside* restricted the domain of reference as a function of which of the objects in the scene could accommodate the previously mentioned object ('*the cube*'). This situation has much in common with the hypothetical one described above, in which the domain of reference could in principle be restricted at a verb such as *eat* as a function of which foodstuff would plausibly be eaten by whoever (or whatever) the previous sentential subject referred to. However, Chambers et al. (1998) did not report the time-course of these effects relative to the onset of the critical referring expression (cf. Figs. 2 and 3 above), and the interaction with definiteness suggests that the critical saccadic movements were programmed during the referring expressions themselves. It is thus difficult to determine on the basis of the available data whether restriction of the domain of reference as a function of the identity of the pre-prepositional direct object took place *in anticipa-*

tion of the post-prepositional referring expression (the effects reported in Experiments 1 and 2 (Sections 2 and 3) (which did anticipate the post-verbal referring expressions).

The Chambers et al. (1998) data speak to the present account because one might suppose that it is not verb-specific knowledge *per se* that is crucial with respect to the projection and referential/anaphoric evaluation of as yet unrealized grammatical objects, but rather the knowledge that is associated, through experience, with *any* theta assigner (that is, with any part of speech that defines roles and/or relationships between entities) – including prepositions. In which case, the Chambers et al. (1998) experiments are very relevant to the issue of whether the contingency between a theta-assigner and a previously mentioned recipient of one of its roles can be used to anticipate, at that theta–assigner, whatever might subsequently be referred to. Evidently, further research is required to further resolve this issue, and to explore further the relationship between the Chambers et al. (1998) findings and our own.

Eberhard et al. (1995) and Sedivy et al. (1999) have shown how the process of reference is time-locked to the accumulation of information within a referring expression. Chambers et al. (1998) have shown that information conveyed by a preposition can restrict the domain of interpretation for the immediately following referring expression. Our demonstration goes one step further, and shows how the process of reference can be time-locked to the accumulation of information received not just during referring expressions themselves, but to the accumulation of information received during the processing of the linguistic entity – the verb – whose interpretation determines the real-world relationships between the entities referred to by those expressions. The data point to a process that is highly predictive, in which any and all available information (subject to empirically testable limits – see above) is recruited to the task of predicting subsequent input. We believe our data address fundamental issues concerning the nature of the representations that are evoked as a sentence unfolds, and how these representations are modulated by the context within which that unfolding occurs. Our data suggest, for example, that the processor can predictively activate representations corresponding to a verb's arguments, and in doing so, evaluate those arguments against the context – thus, the representations that are constructed during sentence processing encode not simply the dependencies, or contingencies, between the current linguistic input and future possible input, but also the dependencies between that future input and the current (or anticipated) context.

Acknowledgements

This work was carried out with the support of grants G9628472N from the Medical Research Council, R000222798 from the Economic and Social Research Council, and RG17935 from The Royal Society. Colin Blackburn developed the software for stimulus presentation and EyeLink control. We thank Holly Branigan for making available to us software developed at the University of Glasgow for equating screen coordinates output by EyeLink with the actual objects contained

within the scene. We also thank Kate Nation for helpful discussions, and Sarah Haywood (who helped run Experiment 2 (Section 3)) and three anonymous reviewers for their comments on a previous version of this article.

Appendix. A

The sixteen sentence pairs used in Experiments 1 and 2 (Sections 2 and 3). The selectional requirements of the first verb in each pair were fulfilled by just one object in the accompanying visual scene, whilst the requirements of the second verb in each pair were fulfilled by more than one object. The objects in parentheses refer to the distractor objects included in the accompanying visual scene. In cases with four distractors, the final distractor in the list was deemed incompatible with (or implausible as the object of) either verb. The visual scenes can be viewed at http://www-users.york.ac.uk/ ~ gtma1/cog99/appendix.html.

1. The boy will eat/move the cake. (toy car, ball, toy train).
2. The woman will drink/try the wine. (cheese, lipstick, chair, plant).
3. The policeman will arrest/search the man. (car, dustbin, houses, cat).
4. The woman will bathe/touch the baby. (plant, rocking-horse, stool).
5. The boy will bounce/throw the ball. (paper plane, shuttle-cock, acorns, bicycle).
6. The hiker will climb/photograph the mountain. (animal, moon, cactus).
7. The housewife will fry/wash the mushrooms. (knife, jug, weighing scales).
8. The doctor will inject/check the child. (TV monitor, microscope, books, toy bear).
9. The woman will play/dust the piano. (table, television, telephone).
10. The woman will read/open the book. (door, bag, jar, cup).
11. The man will repair/wipe the washing machine. (mirror, wastebin, dog).
12. The baby will ring/kick the bell. (drum, bricks, duck).
13. The man will sail/watch the boat. (birds, car, sun).
14. The man will smoke/collect the cigarettes. (filofax, glasses, briefcase, clock).
15. The boy will walk/feed the dog. (bird, pig, hen, ball).
16. The businessman will wear/forget the hat. (wallet, folder, magnifying glass, chair).

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *Journal of Memory and Language*, *38* (4), 419–439.

Altmann, G. T. M., & Steedman, M. J. (1988). Interaction with context during human sentence processing. *Cognition*, *30* (3), 191–238.

Altmann, G. T. M. (1997). *The ascent of Babel: an exploration of language, mind, and understanding*, Oxford: Oxford University Press.

Altmann, G. T. M. (1999). Thematic role assignment in context. *Journal of Memory and Language*, *41*, 124–145.

Bresnan, J. (1982). *The mental representation of grammatical relations*, Cambridge, MA: MIT Press.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Carlson, G. N., & Filip, H. (1998). *Words and worlds: the construction of context for definite reference. Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, Mahwah, N.J: Lawrence Erlbaum pp. 220–225.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6* (1), 84–107.

Eberhard, K., Spivey-Knowlton, S., Sedivy, J., & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language processing in natural contexts. *Journal of Psycholinguistic Research*, *24*, 409–436.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101* (4), 676–703.

Matin, E., Shao, K., & Boff, K. (1993). Saccadic overhead: information procesing time with and without saccades. *Perception and Psychophysics*, *53*, 372–380.

McRae, K., Ferretti, T. R., & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, *12*, 137–176.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109–147.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268* (5217), 1632–1634.