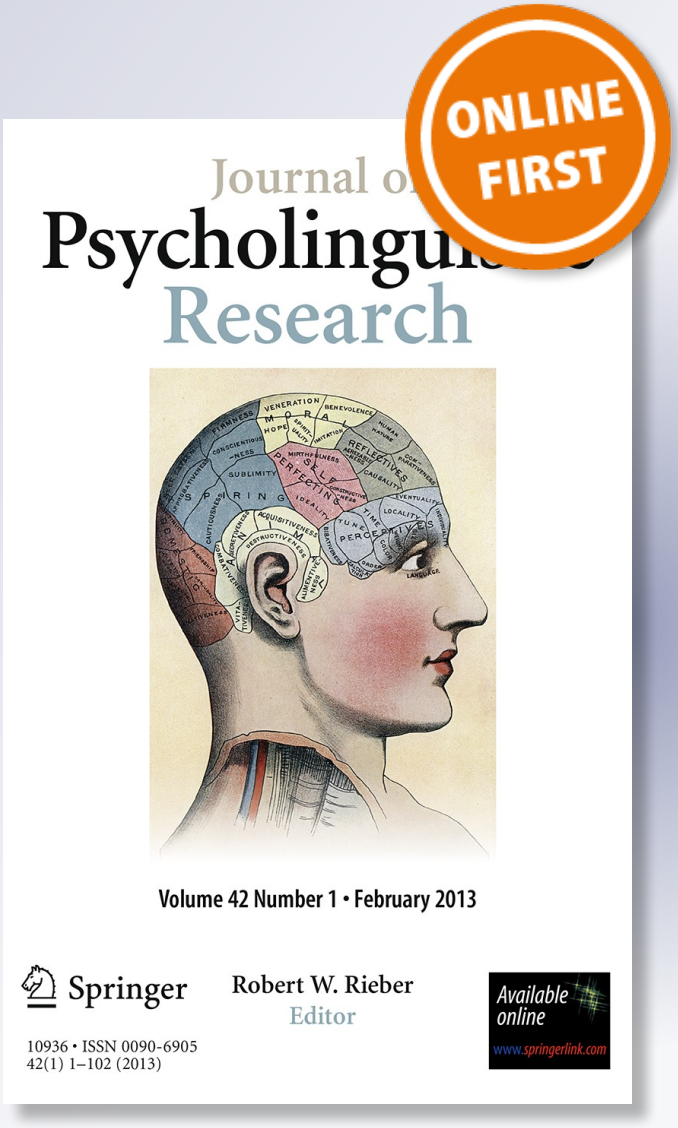
Good-Enough Language Processing: Evidence from Sentence-Video Matching

Gaurav Kharkwal & Karin Stromswold

Journal of Psycholinguistic Research

ISSN 0090-6905

J Psycholinguist Res
DOI 10.1007/s10936-013-9239-5



Springer

J Psycholinguist Res
DOI 10.1007/s10936-013-9239-5

Good-Enough Language Processing: Evidence from Sentence-Video Matching

Gaurav Kharkwal · Karin Stromswold

Abstract This paper investigates how detailed a linguistic representation is formed for descriptions of visual events. In two experiments, participants watched captioned videos and decided whether the captions accurately described the videos. In both experiments, videos depicted geometric shapes moving around the screen. In the first experiment, all of the captions were active sentences, and in the second experiment, half of the captions were active and half were passive. Results of these experiments indicate that participants who only encountered active sentences performed less detailed analyses of the sentences than participants who encountered both active and passive sentences, suggesting that the level of linguistic detail encoded reflects the complexity of the task that participants have to perform. These results are consistent with "good enough" models of language processing in which people process sentences heuristically or syntactically depending on the nature of the task they must perform.

Keywords Good-enough models · Sentence processing · Sentence-video matching · Passive sentences

Introduction

Previous research on how the visual and the linguistic cognitive systems interact has suggested that the two systems are closely integrated and that information presented to one system can

G. Kharkwal · K. Stromswold
Department of Psychology, Rutgers University, New Brunswick, NJ, USA

G. Kharkwal · K. Stromswold
Center for Cognitive Science, Rutgers University, New Brunswick, NJ, USA

G. Kharkwal (✉)
152 Frelinghuysen Road, Piscataway 08854, NJ, USA
e-mail: kharkwal@rutgers.edu

Published online: 06 February 2013

influence the processing of information presented to the other (e.g., Tanenhaus et al. 1995; Sedivy et al. 1999; Altmann and Kamide 1999; Altmann 2004; Knoeferle et al. 2005).

Another interesting aspect of visual-linguistic integration is the nature of verbal descriptions of the visual world. Even a concrete visual scene can be described verbally in many ways. For example, a scene depicting a man and a woman standing together can be described in the following ways: the man can be said to be standing to the left of the woman, the woman can be said to be standing to the right of the man, the man and the woman can be said to be standing next to each other, and so on. Things become even more complex in dynamic visual scenes (henceforth, visual events). Consider a visual event in which two things are moving together, with one being in front of the other. The verbs, *chase*, *flee*, *lead*, *follow*, *trail*, *guide*, etc. might all be used to describe such a visual event. Factors such as the speed (and changes in the speed) of the two objects and the distance between the two objects (and changes in inter-object distance) likely affect what is the "best" verb to use, but there is no set value for any of these factors that unambiguously distinguishes one event from another, and, ultimately, the difference lies in the context.

For example, consider the case of two cars moving such that one car is behind the other. If the two cars move at more or less the same speed and the distance between them stays more or less the same, *leading* or *following* might seem apt descriptions. On the other hand, if the two cars move at high speeds and the distance between them changes often, *chasing* or *fleeing* might seem better.

Irrespective of the speeds of the two cars and the distances between them, the choice of verb changes depending on the perspective from which the event is described. If the event is described from the perspective of the rear car, *chasing*, *following*, *trailing*, etc. can be used. On the other hand, if the event is described from the perspective of the front car, *fleeing*, *leading*, *guiding*, etc. can be used instead. Previous work has suggested that people have a bias towards descriptions in which the subject of the sentence is the "source" of the action and the object is the "goal" (e.g. Fisher et al. 1994; Lakusta and Landau 2005). That is, people tend to describe the same event as either *chasing* or *following* instead of *fleeing* or *leading*.

The choice of verb used can also be influenced by the entities involved in the action (i.e. the nouns). The same event might be better described as *chasing* instead of *following* if the entities involved are more animate, as animacy often entails features like intentionality and aggression, factors which may distinguish *chasing* from *following*. For example, in a similar event involving a dog and a rabbit, the verbal label is more likely to be *chasing* or *fleeing* than *following* or *leading*. Conversely, if the entities were geometric shapes, *following* or *leading* might be better than *chasing* or *fleeing*.

How crucial is the choice of verb used to describe an event, and to what extent does the choice of verb affect the linguistic representations people form when they process language? Researchers disagree as to the nature of the representations that people build when they process sentences, with some arguing that people syntactically parse sentences and create detailed representations (Frazier 1978; MacDonald et al. 1994; Trueswell et al. 1994) and others arguing that that is not the case. For example, Bever and colleagues have argued that people often use non-syntactic heuristics to process sentences. In early work, Bever (1970) argued that people assume that the sentences they hear exhibit a canonical structure (e.g., in English, "Noun Verb Noun") and that the constituents have specific semantic roles (e.g., in English, that the first Noun Phrase (NP) is the agent and the second NP is the patient). More recently, Townsend and Bever (2001) have argued that sentence comprehension is a two-step process. In their Late Assignment of Syntactic Theory (LAST) model, non-syntactic heuristics first extract lexical information and attribute thematic roles to the various constituents and create a "pseudo-syntactic" representation of the sentence. That representation is then used

as input by an algorithmic parser that constructs a final, syntactic representation that is then compared with the input sentence for verification. Thus, if only the first stage of processing occurs (i.e. only a pseudo-syntactic representation is created) before people perform a task, the choice of verb should not affect people's performance. However, if the final, detailed syntactic representation is produced, subtle differences in the meanings of verbs might be represented and the choice of verb might affect people's performance.

In a similar vein, Ferreira and colleagues have hypothesized that the representations created during language comprehension are not necessarily exact and are often simply "good enough" (e.g., Ferreira and Henderson 2007; Ferreira 2003; Ferreira et al. 2002; Christianson et al. 2001). They argue that the details in the final representation depend on the nature of the task that the listener wishes to perform. When the task does not require a detailed representation people use non-syntactic heuristics to create a "quick and dirty" parse, and when the task requires a detailed representation, they use algorithmic parsing. If indeed the nature of the final representation is merely good enough for the task required, then for some tasks the verb used to describe an event might not affect people's performance.

In the two experiments described below, we used the fact that visual events can be described by different verbs to investigate the extent to which adults create detailed linguistic representations when they process sentences. In these experiments, participants watched captioned videos and decided whether the caption accurately described the video. We investigated whether subtle differences in the videos and in the captions affected people's performance. In the first experiment, all of the captions were active sentences and, thus, a heuristic parse is all that is needed to successfully perform the task. In the second experiment, half of the captions were active sentences and half were passives and, thus, heuristic parsing is not sufficient for successful performance.

Experiment 1

Methods

Participants

Twenty-five native, monolingual English-speaking college students participated in the experiment for course credit. All had normal or corrected-to-normal vision, and none had a history of hearing loss or a language or learning disorder.

Stimuli and Apparatus

The stimuli and the experiment were programmed and presented using PyGame (http://www.pygame.org) on a 21 inch flat-screen LCD display with 1,920 × 1,080 pixels resolution. Participants sat approximately 50 cm. away from the screen, and all the visual angle measurements done below are based on that viewing distance.

Visual Stimuli Each trial had a visual and a linguistic component. The visual component was an animated event depicting two geometric shapes moving within the confines of a bounding box, with one shape always being behind the other (See Fig. 1). The horizontal side of the bounding box subtended a visual angle of 37° and the vertical side subtended an angle of 21°. Four shapes were used (circles, squares, ovals and rectangles) and each scene had two

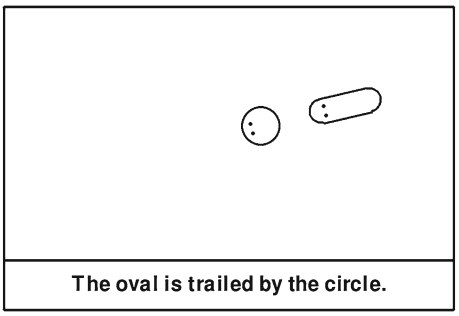


Fig. 1 A screenshot of the display (black-white inverted)

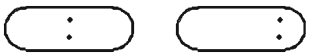


Fig. 2 The no-eyes (*left*) and the eyes (*right*) conditions

different shapes. The circle subtended a visual angle of 1.2°, and the square subtended a visual angle of 1.0°. The rectangle subtended 1.7° along the longer axis and 0.9° along the smaller one, and the oval subtended 2.0° along the longer axis and 0.7° along the smaller one. The elongated shapes (oval and rectangle) were rendered such that the longer axis was parallel to the direction of motion.

The shapes' initial positions and headings were randomly generated before the start of each trial, and the leading shape randomly moved away from the shape following it. Thus, every trial had a different overall display. The shapes moved with an average speed of 0.2°/s. As discussed above, intuitively, what verb is best for describing a visual event depends in part on the animacy of the entities involved (the NPs). To investigate whether this is true, a variable controlled the presence or absence of "eyes" on these shapes, with the expectation that shapes with "eyes" would appear more animate. As depicted in Fig. 2, each shape had two dots that were either at the "front" of the object ("Eyes" condition) or the center of the shapes ("No Eyes" condition). The two dots were placed such that the line joining the two lay perpendicular to the axis of motion in order to ensure that as the shapes rotated, so did the two dots. The two dots were separated by a visual angle of 0.3°.

Linguistic Stimuli As shown in Fig. 1, below the bounding box of the animated event was a smaller rectangular box that contained a sentence describing the event (henceforth, the caption). Captions were centered and displayed in 40 pt. font size and subtended visual angles between 11° and 14°. For ongoing actions, the present progressive tense (e.g. *The oval is following the square*) is more felicitous than the simple present tense (e.g. *The oval follows the square*), the simple past tense (e.g. *The oval followed the square*), or the progressive past tense (e.g. *The oval was following the square*). Thus, the syntactic structure for all the captions was: *The* SHAPE_1 *is* VERB*ing the* SHAPE_2. Four verbs were used: *chase*, *flee*, *lead*, and *follow*. Notice that all four verbs can be used to describe a visual event involving two entities such that one is moving behind the other. Two of the four verbs describe the event from the perspective of the shape that is behind (*chase* and *follow*), and the other two describe the event from the perspective of the shape in front (*flee* and *lead*).

In half of the trials, the propositional content of the caption and the visual event matched ('match' trials), and in half of the trials the propositional contents did not match ('mismatch' trials) because the semantic roles of the shapes were switched. For example, in the trial

depicted in Fig. 1, a match caption was, *The oval is following the square*, and a mismatch caption was, *The square is following the oval*.

Design

Each of the 12 shape pairs appeared equally often with the 2 eyes conditions, 4 verbs, and 2 match/mismatch conditions to yield 192 unique trial types. The list of the 192 trial types was pseudo-randomized with the constraint being that no more than 4 consecutive trials contained the same value for any of the three independent variables. Half of the participants received the trials in this order and half received the trials in the reverse order.

Procedure

Participants began each trial by fixating on a crosshair at the center of the screen. When they were ready to begin a trial, they pressed the spacebar at which point the caption and the animated video appeared on the screen simultaneously. Participants were instructed to press the left shift key if the caption matched the video and the right shift key if the caption did not match the video. Response times (RTs) were measured from the moment the spacebar was pressed until the participant hit a shift key. Participants were told to respond as quickly as they could without sacrificing accuracy.

Before the experimental trials, participants did 8 practice trials that were selected such that the value of each independent variable occurred equally often over the course of the practice trials. Participants who made more than one mistake during the practice phase repeated the practice trials until they made no mistakes.

Analysis

RTs for correct trials were analyzed using multi-way ANOVAs with Subject as a random variable. To confirm and measure the strengths of the results obtained from ANOVAs, Bayesian analyses were performed using the methods described by Masson (2011). For the Bayesian analyses, Bayes Factors are given as the ratio of probability of obtaining the observed data given the null hypothesis over the probability of obtaining the observed data given the alternate hypothesis. In other words, the Bayes Factors reported here are odds *favoring* the null hypothesis given the data. The Bayes Factors were estimated using the Bayesian Information Criterion (Raftery 1995). In addition, the posterior probability corresponding to the hypothesis that the data favored are reported. Raftery (1995)'s thresholds were used to categorize the strength of evidence. If the evidence in favor of one hypothesis (i.e. the value $p_{BIC}(H_0|D)$ or $p_{BIC}(H_1|D)$) was between 0.50 and 0.75, it was classified as "weak" evidence; if it was between 0.75 and 0.95, it was classified as "positive" evidence; if it was between 0.95 and 0.99, it was classified as "strong" evidence; and if it was greater than 0.99, it was classified as "very strong" evidence.

Results

Collapsing across all participants' data, participants correctly responded to about 95 % of trials (4,556 out of 4,800). When all trials were included, the mean RT was 2,966 ms (SE = 33.3 ms), and when only correct trials were included, the mean RT was 2,912 ms (SE = 30.9 ms), suggesting that there was no speed-accuracy tradeoff.

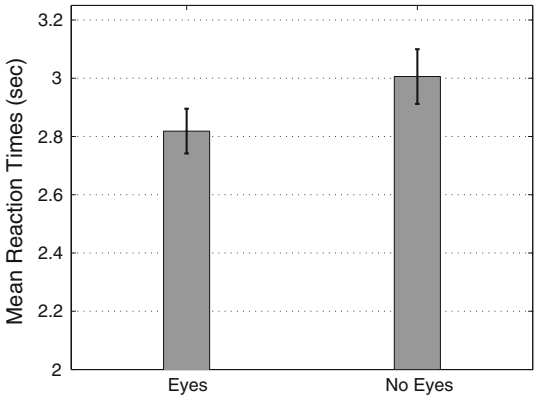


Fig. 3 Effect of eyes on RTs (*error bars* = SE)

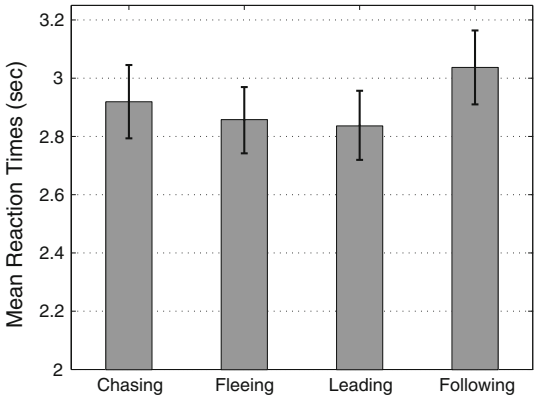


Fig. 4 Effect of verb on RTs (*error bars* = SE)

A 2 (eyes) × 2 (match/mismatch) × 4 (verbs) ANOVA of correct trial RTs revealed a main effect of Eyes with participants responding about 6% faster when the shapes had 'eyes' than when they did not (2,818 and 3,005 ms, respectively; $F(1,24) = 8.17$, $p = 0.009$; Fig. 3). Bayesian analysis provided positive evidence confirming the result ($BF = 0.13$, $p_{BIC}(H_1|D) = 0.89$).

There was no significant effect of verb choice ($F(3,72) = 1.60$, $p = 0.2$), and Bayesian analysis provided strong evidence confirming the result ($BF = 58.15$, $p_{BIC}(H_0|D) = 0.98$; Fig. 4). There was also no significant difference between the match and the mismatch conditions ($F(1,24) = 0.39$, $p = 0.54$; Fig. 5), and Bayesian analysis provided positive evidence confirming the result ($BF = 4.10$, $p_{BIC}(H_0|D) = 0.80$). Lastly, there were no significant interactions for any independent variables, and Bayesian analyses confirmed these results.

To test whether verb perspective played a role, the four verbs were grouped into two pairs based on whether they were "source-to-goal" verbs (*chase* and *follow*) or "goal-to-source" (*flee* and *lead*). As shown in Fig. 6, a 2 (eyes) × 2 (match/mismatch) × 2 (verb perspective) ANOVA of correct trial RTs revealed a significant effect of verb perspective with participants responding about 4% faster for goal-to-source verbs than source-to-goal verbs (2,847 and 2,977 ms, respectively; $F(1,24) = 7.35$, $p = 0.01$). Bayesian analysis provided positive evidence confirming the result ($BF = 0.17$, $p_{BIC}(H_1|D) = 0.85$). Verb perspective did not interact with the other variables, and Bayesian analysis confirmed these results.

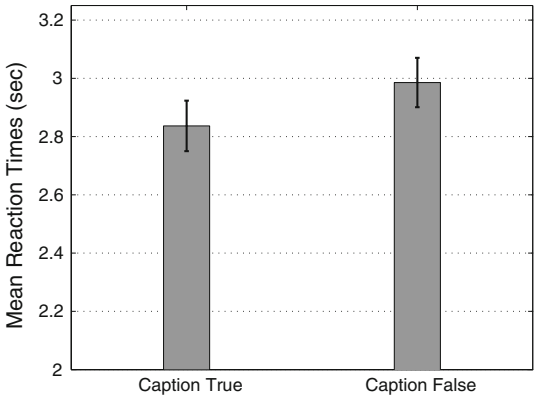


Fig. 5 Effect of match/mismatch on RTs (*error bars* = SE)

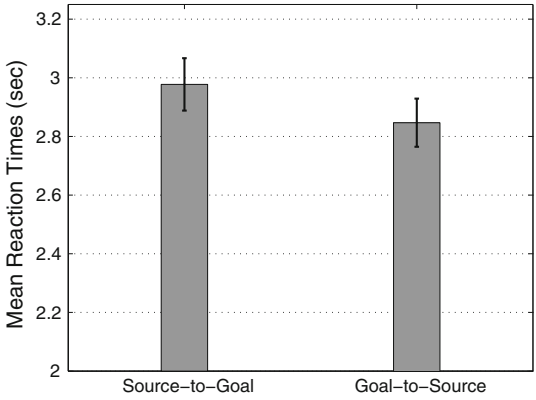


Fig. 6 Effect of verb perspective on RTs (*error bars* = SE)

Discussion

Recall that the shapes moved in exactly the same way in the Eyes trials and the No Eyes trials and the Noun Phrases used to describe these shapes were the same in the Eyes and No Eyes trials. The fact that people were faster on the Eyes trials suggests that the two eccentric dots provided some cue that aided perception of the visual event. One possibility is that these eccentric dots were indeed perceived as 'eyes' leading participants to attribute a certain degree of animacy to these geometric shapes, and, perhaps, perceiving an event as animate makes it easier to perceive, encode, or interpret the event. Most research on animacy perception has investigated visual features that trigger animacy (e.g., Dittrich and Lea 1994; Gelman et al. 1995; Tremoulet and Feldman 2000, 2006; Gao et al. 2009), with animacy perception being an end result of visual processing. Given the nature of our task, it seems unlikely that the visual scene was processed, animacy obtained, and then the thus-obtained animacy information was used to revise the initial representation. Furthermore, if 'eyes' serve as a cue for animacy and animacy of NPs affects verb choice, one could argue that presence of 'eyes' should improve performance for trials captioned with *chase* and *flee* and hinder performance for trials captioned with *lead* and *follow*. The fact that no interaction was found between verb choice and Eyes argues against animacy being the cause of the Eyes effect.

A second, more plausible, explanation for the Eyes effect is that 'eyes' conveyed information about the direction of motion in the visual event. Participants could have used 'eyes'

to infer the direction of motion to determine which shape was in front. They could then have used this information to attribute roles to the two shapes. Perhaps the reason why participants' RTs were greater when 'eyes' were absent (i.e. the two dots were centric) was because the still image was ambiguous and multiple images were required to determine the direction of motion and subsequently process the event.

The lack of a verb effect might reflect that final linguistic representations are less detailed than the original captions (i.e. they are only "good enough"), perhaps just encoding the details of which entity was where. However, note that in order to perform accurately, participants must have encoded the perspective of the verb. For example, in *The square is chasing the oval*, the first NP corresponds to the trailing shape and the second to the leading shape. On the other hand, in *The square is fleeing the oval*, the first NP corresponds to the leading shape and the second to the trailing shape. Thus, to perform accurately participants needed to know which NPs corresponded to the trailing and the leading shapes, information which they could only get after encoding the perspective of the verb. Consistent with that account, we did find an effect of verb perspective.

Previous work (e.g., Fisher et al. 1994; Lakusta and Landau 2005) suggests that people have a semantic bias for source-to-goal verbs over goal-to-source verbs. One way to reconcile this with our finding that participants were faster for goal-to-source verbs than source-to-goal verbs is to argue that, in our study, goal-to-source verbs are better descriptors for the events depicted in our videos, although pretesting demonstrated that the videos could be described using any of the four verbs. One of the subtleties of our visual display is that near the corners, the shape that is behind slows down and allows the shape in front to move away. We purposely did this to ensure that the trailing shape never "catches up" with the leading shape. It may be that this makes it seem as if the trailing shape lets the leading shape "guides" or "leads" it.

Another possibility is that previous studies that have revealed semantic bias in favor of source-to-goal verbs have been production studies in which children describe a visual scene, whereas our study is a comprehension study involving adults. Indeed, pilot data indicate that when adult participants are asked to describe the visual events in our videos, they are more likely to use source-to-goal verbs than goal-to-source verbs. Production and comprehension are complementary processes, with the former involving mapping from conceptual structures to linguistic elements, and the latter involving mapping a linguistic input to conceptual structures. Our findings that participants perform faster when goal-to-source verbs are used coupled with the results of our pilot study suggest that there is not a one-to-one mapping between conceptual structures and linguistic representations. That is, even when a concept is best described in a particular way, it does not necessarily imply that the listener will find that description optimal. It may be that other factors, e.g., frequency, might influence what is considered an optimal description. As a case in point, the source-to-goal verbs (*chase*, *follow*, *trail*, *pursue*, etc.) occur more frequently in the Kucera and Francis (1967) database than the goal-to-source verbs (*flee*, *lead*, *guide*, *evade*, etc.)

Results of many different types of cognitive experiments (including some sentence-picture verification studies) suggest that people take less time to decide that something is true than to decide that something is *not* true. In sentence-picture verification studies, the greater RTs for mismatch (i.e. false trials) than match trials (i.e. true trials) is generally attributed to the cost of verifying a mismatch (Carpenter and Just 1975; Clark and Chase 1972). That is, in case of a mismatch, the system restarts the process of comparison resulting in a greater, overall cost. Consistent with the results of other sentence-picture verification studies that have *not* found an effect of match/mismatch (e.g., Underwood et al. 2004; Knoeferle and Crocker 2005), participants in Experiment 1 were no faster on match trials than mismatch trials. One possible explanation for this is that participants processed the sentences so quickly (perhaps using

non-syntactic heuristics) that the subsequent mismatch verification phase was fast enough to not be apparent in overall sentence RTs. This explanation is consistent with Knoeferle and Crocker (2005)'s finding that total sentence reading times often fail to find a match effect that is detectable when fine-grained, constituent-based analyses are performed.

In Experiment 1, all of the captions were active sentences in which the first NP was the agent of the sentence. As a result, participants could have correctly interpreted the sentences by merely using a simple non-syntactic heuristic such as Bever (1970)'s N(oun) V(erb) N(oun) = "Agent Action Patient" heuristic. The results of Experiment 1 can also be explained by the LAST model by Townsend and Bever (2001). Participants could have been using the pseudo-syntactic representations derived after the first phase of the process of comprehension. For active sentences, a pseudo-syntactic parse is similar to the output of an N V N heuristic, with the first Noun Phrase being assigned the agent role and the second the patient.

However, the N V N template or the pseudo-syntactic parse might not work for different kinds of sentences. For example, in English passive sentences, the mapping between grammatical and thematic roles is switched. That is, the subject of a passive sentence is the patient and the object is the agent. Using the N V N template for a passive sentence, '*the square is chased by the oval*,' results in an incorrect interpretation where the square is chasing the circle. Therefore, if the N V N template is used to parse the sentences, all passive sentences would be incorrectly parsed.

On the other hand, the LAST model handles passive sentences by utilizing lexical information in passivized verbs. The information signals to the processor that the sentence is a passive sentence and hence the initially incorrect parse needs to be revised. This process of revision takes place during the first phase itself and results in a second pseudo-syntactic encoding. Thus, for passive sentences, the LAST model offers two different predictions. Participants could either continue relying on pseudo-syntactic representations or use the fully-formed, syntactic representation instead. If participants continue to rely on pseudo-syntactic representations for the task, we expect to see no differences between the verbs. However, if syntactic representations are used, we might see an effect of verb choice. Because the model revises the initial model during the first phase, both possibilities predict no difference in accuracy between actives and passives. Also, because the system needs to revise its initial incorrect parse in the case of passive sentences, both possibilities predict a difference in reading time performance between actives and passives.

In order to test these predictions, we conducted a second experiment, where the visual display was the same as the first experiment but included both active and passive sentences.

Experiment 2

Methods

Participants

Twenty-four native, monolingual English-speaking college students participated in the experiment for course credit. All had normal or corrected-to-normal vision, and none had a history of hearing loss or a language or learning disorder.

Stimuli and Apparatus

The monitor and the visual component of the trials were the same as in the first experiment.

The linguistic stimuli differed from the first experiment in two ways. First, whereas all of the captions in Experiment 1 were active sentences, in Experiment 2, half of the captions were active sentences and half were passive sentences. The active sentences were again in a fixed, present progressive form (i.e. *The* SHAPE_1 *is* VERB*ing the* SHAPE_2). The passive sentences were also of a fixed form and had the following syntactic structure: *The* SHAPE_2 *is* VERB*ed by the* SHAPE_1. Passive sentences like *The oval is chased by the square* differ from active sentences like *The square is chasing the oval* in four ways. First, The semantic roles are reversed in the passive form. Second, although both sentences contain the auxiliary word *is*, in the active sentence, the word is an active progressive auxiliary, and in the passive, it is the homophonous passive auxiliary. The passive auxiliary is different from a passive progressive and we can have a sentence containing both, for example, *The oval is being chased by the square.* Third, the passive sentences have verbs with a passive participle (*-ed*), whereas the active sentences have verbs with the progressive participle (*-ing*). Lastly, the passive sentences contain the preposition *by*.

Additionally, because *flee* does not passivize (**the oval is fled by the square*), we replaced both *flee* and its semantic pair *chase* with *guide* and *trail*. Notice that the replacement verbs *guide* and *trail* also, like *chase* and *flee*, describe visual events where two entities are moving and one is behind the other, albeit from two different perspectives.

Design

Each of the 12 shape pairs appeared equally often with the 2 eyes conditions, 4 verbs, 2 match/mismatch conditions, and 2 syntactic voices to yield 384 unique trial types. The list of the 384 trial types was pseudo-randomized with the constraint being that no more than 5 consecutive trials contained the same value for any of the four independent variables. Half of the participants received the trials in this order and half received the trials in the reverse order.

Procedure

The experimental procedure was the same as in Experiment 1.

Analysis

Data were analyzed in the same way as in Experiment 1.

Results

Collapsing across all participants' data, participants correctly responded to about 96 % of trials (8,824 out of 9,216), with they being equally accurate on actives and passives. When all trials were included, the mean RT was 3,811 ms (SE = 30.8 ms), and when only correct trials were included, the mean RT was 3,836 ms (SE = 31.8 ms), suggesting that there was no speed-accuracy tradeoff.

A 2 (eyes) × 2 (match vs. no match) × 4 (verbs) × 2 (syntactic voice) ANOVA of correct trial RTs failed to reveal a significant effect of Eyes ($F(1,23) = 1.19$, $p = 0.29$; Fig. 7), and Bayesian analysis provided weak evidence in support of the hypothesis that there was no effect ($BF = 2.67$, $p_{BIC}(H_0|D) = 0.73$).

There was a significant effect of verb choice with participants responding fastest when the verb was *lead*, followed by *guide*, and taking approximately the same amount of time for

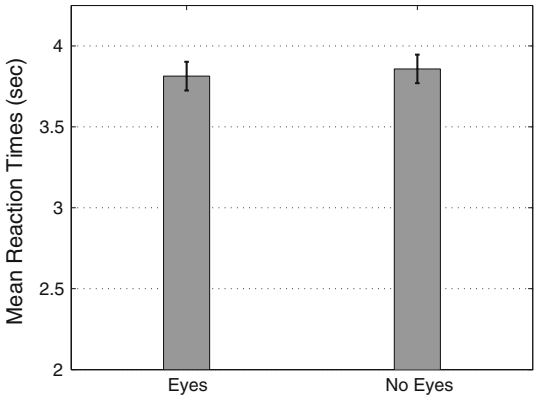


Fig. 7 Effect of eyes on RTs (*error bars* = SE)

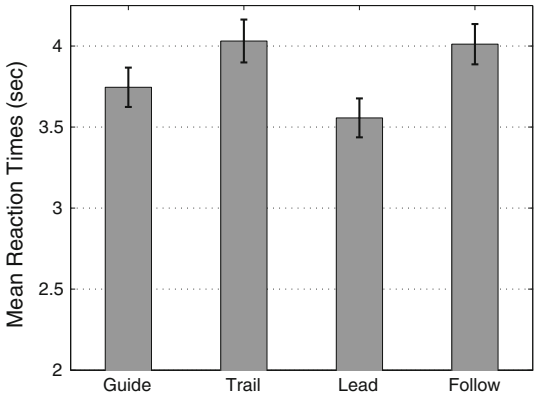


Fig. 8 Effect of verb on RTs (*error bars* = SE)

follow and *trail* (*lead* = 3,556 ms, *guide* = 3,745 ms, *follow* = 4,012 ms, and *trail* = 4,031 ms; $F(3,69) = 12.21$, $p < 0.001$; Fig. 8). Bayesian analysis provided very strong evidence confirming the result ($BF = 0.0001$, $p_{BIC}(H_1|D) = 0.999$).

There was also a significant difference between the match and mismatch conditions with participants responding about 9 % faster when the caption matched the video than when it did not (3,650 and 4,021 ms, respectively; $F(1,23) = 15.73$, $p = 0.001$; Fig. 9). Bayesian analysis provided very strong evidence confirming the result ($BF = 0.0094$, $p_{BIC}(H_1|D) = 0.991$).

Participants were about 12 % faster on active sentences than passive sentences (3,587 and 4,088 ms, respectively; $F(1,23) = 29.87$, $p < 0.001$; see Fig. 10). Bayesian analysis provided very strong evidence confirming the result ($BF = 0.0002$, $p_{BIC}(H_1|D) = 0.999$).

As depicted in Fig. 11, there was a significant interaction between verb choice and match conditions ($F(3,69) = 4.75$, $p = 0.005$) and Bayesian analysis provided weak evidence supporting the interaction ($BF = 0.71$, $p_{BIC}(H_1|D) = 0.58$). Inspection of Fig. 11 suggests that the interaction is due to the verb *guide*, and a post-hoc ANOVA revealed than when data from *guide* trials were excluded, the interaction was no longer significant ($F(2,46) = 1.53$, $p = 0.23$; $BF = 10.23$, $p_{BIC}(H_0|D) = 0.91$). There were no other significant interactions, and Bayesian analyses confirmed these results.

As in Experiment 1, we grouped the verbs as "source-to-goal" (*trail* and *follow*) or "goal-to-source" (*guide* and *lead*) verbs. A 2 (eyes) × 2 (match vs. no match) × 2 (verb perspective) × 2 (syntactic voice) ANOVA of correct trial RTs revealed a significant effect of verb

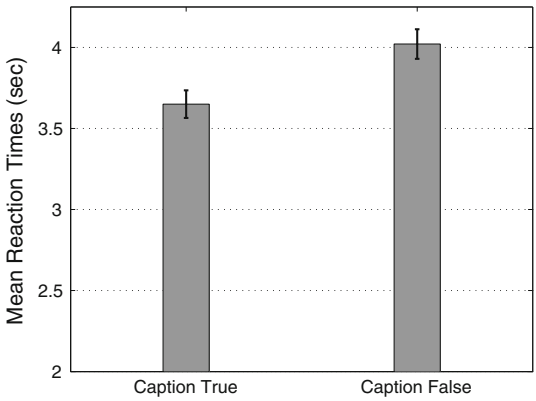


Fig. 9 Effect of match/mismatch on RTs (*error bars* = SE)

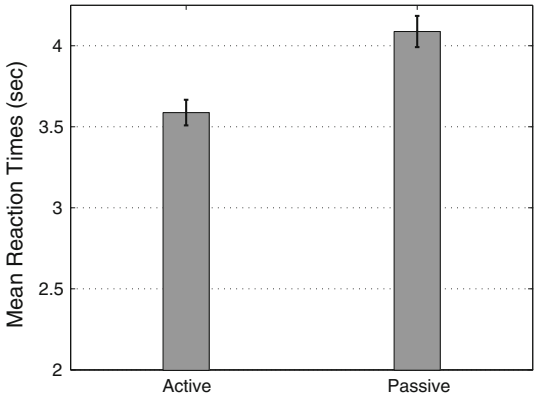


Fig. 10 Effect of syntactic voice on RTs (*error bars* = SE)

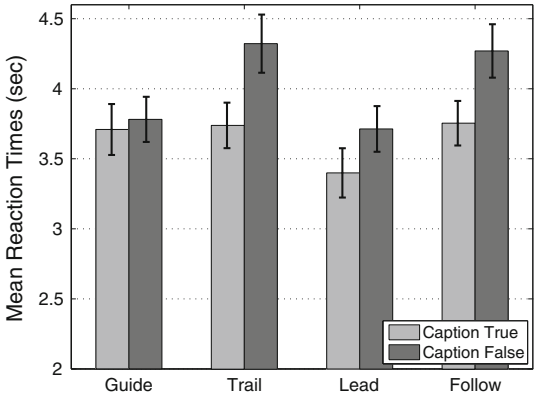


Fig. 11 Interaction between verb and match/mismatch (*error bars* = SE)

perspective with participants responding about 9 % faster when the verbs belonged to the goal-to-source group than when they belonged to the source-to-goal group (3,651 and 4,021 ms, respectively; $F(1,23) = 25.82$, $p < 0.001$; Fig. 12). Bayesian analysis provided very strong evidence confirming the result ($BF = 0.0006$, $p_{BIC}(H_1|D) = 0.999$).

As shown in Fig. 13, there was a significant interaction between verb perspective and the match conditions ($F(1,23) = 11.22$, $p = 0.003$). Bayesian analysis confirmed the result and

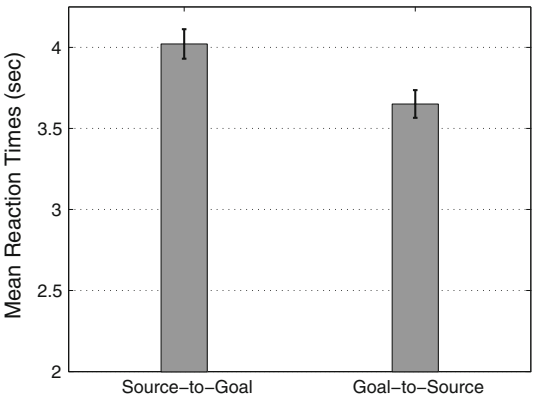


Fig. 12 Effect of verb perspective on RTs (*error bars* = SE)

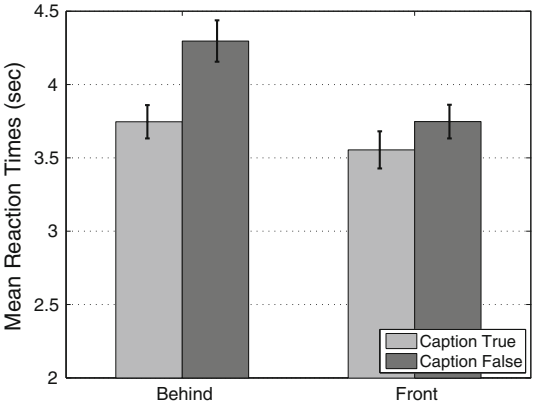


Fig. 13 Interaction between verb perspective and match/mismatch (*Error bars* = SE)

provided strong evidence supporting the interaction ($BF = 0.04$, $p_{BIC}(H_1|D) = 0.96$). Inspection of Fig. 13 suggests that the interaction is the result of a greater "cost" in mismatch trials for source-to-goal verbs, with the differences between match RT and mismatch RT being 550 ms for source-to-goal verbs and 194 ms for goal-to-source verbs. In a post-hoc analysis, we segregated the data by verb perspective and found that Bayesian analysis provided very strong evidence favoring the effect of caption veracity when only source-to-goal verbs were considered ($F(1,23) = 16.44$, $p < 0.001$; $BF = 0.008$, $p_{BIC}(H_1|D) = 0.992$). In contrast, the strength of the evidence favoring the effect was relatively less strong when only goal-to-source verbs were considered ($F(1,23) = 7.46$, $p = 0.012$; $BF = 0.17$, $p_{BIC}(H_1|D) = 0.86$). However, it may be that the apparent difference between source-to-goal and goal-to-source verbs is driven by *guide*. There were no other significant interactions, and Bayesian analyses confirmed these results.

Discussion

To a first approximation, the results of the second experiment are complementary to the results of the first: In the first experiment, only the visual parameter, the presence or absence of 'eyes,' and one linguistic parameter, the perspective of the verb, had an effect, and in the second experiment, everything but the visual parameter had an effect. There are two possible explanations for Eyes not having an effect in Experiment 2. First, participants might not have

been affected by the presence or absence of 'eyes' on the shapes, but this seems unlikely given that the visual stimuli were identical to those used in Experiment 1. Second, perhaps Eyes did have an effect in Experiment 2, but the effect was too transient to be detected in end-of-trial RTs. The second possibility is consistent with Bayesian analysis providing only weak evidence against the effect of Eyes.

Our finding that the linguistic variables, verb choice and syntactic voice, played a role in Experiment 2 but not in Experiment 1 is consistent with participants having fully parsed sentences in Experiment 2, but not in Experiment 1. The greater RTs for passives than actives may reflect the processing cost of revising the initial incorrect representation for passive sentences that resulted from a heuristic-based, pseudo-parse. Our finding that subjects were slower on some verbs than others may reflect a differential cost of comparison of the visual representation with a detailed, syntactic representation of the sentences. The difference between the costs of comparison for the four verbs may well reflect the difference between the verbs as descriptive labels of the visual event, with the better descriptors resulting in smaller cost and faster performance. In Experiment 1, where we hypothesized that participants only used a pseudo-syntactic representation for the task, the very subtle differences in the meanings of the verbs might not have been encoded. As a result, comparisons between the visual and the linguistic descriptions would not reflect inter-verb differences.

Furthermore, the fact that in Experiment 2 (but not Experiment 1) verbs interacted with the match conditions also suggests that, in Experiment 2, the four verbs were processed differently. Our finding that the cost of verification in case of mismatch trials was not as high for goal-to-source verbs as for source-to-goal verbs suggests that descriptions involving goal-to-source verbs may be closer to the visual representation of the event, as a result of which, verifying a mismatch might have been easier.

Because a pseudo-syntactic representation might not encode as many details as a detailed, syntactic structure, we expected verification in case of a mismatch to have a greater cost when a detailed representation is formed. Our finding that participants in Experiment 2 were slower for trials in which the verbal caption and the video did not match is consistent with that prediction, and further suggests that participants syntactically parsed the sentences.

General Discussion

The fact that participants were 32 % slower in Experiment 2 suggests that people may have processed sentences differently in the two experiments. The slower speed in Experiment 2 cannot be the result of a speed-accuracy trade-off as participants were equally accurate in the two experiments. Nor can the slower speed in Experiment 2 be the result of averaging RTs of actives and passives because participants were 23 % slower on actives in Experiment 2 than Experiment 1. It also cannot be that participants in Experiment 1, but not in Experiment 2, just parsed the first noun and then matched the noun with the trailing (or the leading) shape, because the inclusion of perspective shift verbs would have resulted in 50 % accuracy.

In Experiment 1, because all sentences had a canonical NVN structures, participants could have done a "rough" parse. We believe that the resulting representation is a basic, pseudo-syntactic representation that encodes thematic roles and a "who-goes-where" description of the two shapes. As a result, participants performed similarly on all four verbs. Because of inclusion of passives in Experiment 2, a similar strategy is inadequate and a more detailed analysis in necessary. We believe that the resulting representation is a detailed representation that encodes all the information that makes one verb different from another, and as a result, participants performed differently for the four verbs.

A sentence-video matching task cannot be performed until both the sentence and the video have been processed. In Experiment 2, but not in Experiment 1, a detailed analysis may have resulted in sentence processing being the limiting factor, thus overshadowing the effects of visual parameters. Consistent with this account, in Experiment 1, both the visual parameter (Eyes) and the slightly broader linguistic parameter (verb perspective) had significant effects, while in Experiment 2, only the linguistic parameters (voice, verb, verb perspective, and caption veracity) had significant effects.

Our results are consistent with Ferreira and colleagues' hypothesis that the mechanisms in language processing are only "good enough" for the task at hand (Ferreira and Henderson 2007; Ferreira 2003; Ferreira et al. 2002; Christianson et al. 2001). An explicit model of language comprehension that also employs both heuristics and a syntactic parser is Townsend and Bever (2001)'s LAST model. As discussed earlier, their model suggests that language is comprehended in two phases. In the first phase, heuristics are used to generate a pseudo-syntactic parse, which is then fed as input to a syntactic parser that generates a complete parse. Our results are consistent with the LAST model. In Experiment 1, participants may have used the result of the first phase to do the task, and in Experiment 2, participants may have used the output of the syntactic parser.

However, our results do not necessitate a two-phase comprehension model. Another possibility is that the comprehension system simultaneously parses the input using a heuristic and an algorithmic approach. Furthermore, it may also be that there are more than one heuristic parsers running simultaneously along with an algorithmic parser. For example, one, very primitive, parser could use the canonical template proposed by Bever (1970) that assigns the first noun phrase (NP) the agent role and the second NP the patient/theme role. Another heuristic parser could be smarter and may use lexical cues to revise its initial guess, similar to the pseudo-syntactic parser of the LAST model. Thus, even though our results point to a model of comprehension that assumes both a heuristic and an algorithmic approach to language processing, they do not specify how the two approaches are integrated.

Another question that is left unanswered is how the comprehension system decides which of the two (or more) parses of the input to use for a task. One possibility is that the system uses local, sentence-level information to distinguish between 'simple' and 'complex' sentences, where 'simple' sentences are those that can be parsed using non-syntactic heuristics. For example, the system could use information embedded in the meaning of a passivized verb to detect the presence of a passive sentence (Townsend and Bever 2001). Alternatively, the system could use structural features to distinguish between sentences. For example, as discussed earlier, passive sentences are structurally different from active sentences. In our study, either the presence of the '*-ed*' suffix on the verb or the word '*by*' or the combination of the two could have been used as a cue to detect passive sentences. Either ways, on identifying a 'complex' sentence, the system might prefer the output of an algorithmic parser over a heuristic parser, simply because the former is more likely to be accurate.

The second, and perhaps more likely, possibility is that the system uses situational cues to decide which parser to use. For example, the system might start with a "default" behavior always using either the output of the heuristic parser or the algorithmic parser. If the default is the heuristic parser, the system would always make an error when a sentence does not conform to the canonical template. The system could then either switch to the algorithmic parser immediately, or after the number of errors has crossed a certain threshold. Alternatively, if the default is the algorithmic parser, the system could keep a copy of the pseudo-parse and simultaneously use both representations. If the pseudo-parse appears to be sufficient (based on some threshold), the system could then start using the heuristic parser instead.

In conclusion, the results of the experiments presented here suggest that the human language comprehension system employs both heuristic and algorithmic methods to process sentences, choosing the output of one over the other based on task details. The general implication is that linguistic representations are not always comprehensive and may often be merely good enough.

Acknowledgments This research was supported by awards from the National Science Foundation (BCS-0446850 and BCS-0124095) to K.S. We thank the audience at the 11th Annual Meeting of the Vision Sciences Society and the 25th Annual CUNY Conference on Human Sentence Processing for helpful discussions on presentations of this work. We especially thank Jacob Feldman, Eileen Kowler, Nora Isacoff, Choonkyu Lee, Sabrina Angelini, Tina Hou-Imerman, Heather Yaden, and Nikhita Karki for their suggestions and advice in all aspects of this work: from designing the stimuli, to writing this paper.

References

Altmann, G. T. M. (2004). Language-mediated eye-movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition*, *93*, B79–B87.

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

Bever, T. G. (1970). *The cognitive basis for linguistic structures*. New York: Wiley.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*, 45–73.

Christianson, K., Hollingworth, A., Halliwell, J., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*, 368–407.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517.

Dittrich, W., & Lea, S. (1994). Visual perception of intentional motion. *Perception*, *23*, 253–268.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164–203.

Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11–15.

Ferreira, F., & Henderson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*, 71–83.

Fisher, C., Hall, D. G., Rakowitz, S., & Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, *92*, 333–375.

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies*. Unpublished doctoral dissertation, University of Connecticut, Storrs.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psycholphysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, *59*, 154–179.

Gelman, R., Durgin, F., & Kaufman, L. (1995). *Distinguishing between animates and inanimates: Not by motion alone*. Oxford: Clarendon Press.

Knoeferle, P., & Crocker, M. (2005). Incremental effects of mismatch during picture-sentence integration: Evidence from eye-tracking. In *Proceedings of the 26th annual conference of the cognitive science society* (pp. 1166–1171), Stresa, Italy.

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

Kucera, N., & Francis, W. N. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Lakusta, L., & Landau, B. (2005). The importance of goals in spatial language. *Cognition*, *96*, 1–33.

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

Masson, M. E. J. (2011). A tutorial on a practical bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690.

Raftery, A. E. (1995). *Bayesian model selection in social research* (pp. 111–196). Cambridge, MA: Blackwell.

Sedivy, J., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109–148.

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

Townsend, D., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. Cambridge, Ma: MIT Press.

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*, 943–951.

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception and Psychophysics*, *68*, 1047–1058.

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic disambiguation. *Journal of Memory and Language*, *33*, 285–318.

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *The Quarterly Journal of Experimental Psychology*, *56*, 165–182.