LINGUISTIC REPRESENTATIONS OF VISUAL EVENTS

by

GAURAV KHARKWAL

A thesis submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Psychology

Written under the direction of

Dr. Karin Stromswold

and approved by

_____

_____

_____

New Brunswick, New Jersey

October, 2011

ABSTRACT OF THE THESIS

Linguistic Representations of Visual Events

By GAURAV KHARKWAL

Thesis Director:

Dr. Karin Stromswold

This thesis explores the nature of linguistic representations that correspond to verbal descriptions of events. In two experiments, participants watched captioned videos and decided whether the captions accurately described the videos. In the videos, two geometric shapes moved around the screen. [In half of the trials, the geometric shapes had "eyes."] The verbs used to describe the shapes' actions were either source-to-goal verbs (*chase*, *follow*, *trail*) or goal-to-source verbs (*flee*, *lead*, *guide*). Sometimes the captions were active sentences (e.g., *The circle is chasing the square*) and sometimes passive sentences (*The square is chased by the circle*). Analyses of participants' reaction times indicate that the level of linguistic and visual detail encoded reflected the complexity of the task participants had to perform. These results are consistent with "good enough" models of language processing (e.g., Ferreira and Henderson (2007)) in which people process sentences heuristically or syntactically depending on the nature of the task they must perform.

Acknowledgements

First and foremost, I would like to thank Karin Stromswold. Without her there would not be this thesis. She has been a great mentor and has always provided me with invaluable academic and personal advice. I would like to extend my thanks and sincere appreciation to my committee members, Jacob Feldman and Eileen Kowler, for helpful suggestions and guidance, and colleagues in the Language Acquisition and Processing Lab and the Department of Psychology for encouragement. I would also like to thank my family for being supportive of my choices and decisions. Last, but definitely not the least, I thank Nikhita Karki for always believing in me and never letting me give up.

Table of Contents

List of Figures

1. Introduction

We often use language to discuss aspects of the visual world and the listener is tasked with integrating the visual and the linguistic information. 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 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)). In other words, what we see may influence how we interpret an utterance, and, conversely, what we are told may influence the way we inspect the visual world.

Another interesting aspect of visual-linguistic integration is the nature of the verbal descriptions of the visual world. For example, if you wanted to describe a visual scene to a friend, how detailed should your description be? What if you had to describe it in the best possible way? Is there always one "really good" way to describe a visual scene? As it turns out the answer to that last question is no.

Even a concrete visual scene can usually be described verbally in many ways. Take the case of a glass that contains water, is it best described as being half empty or half full? If a man and woman are standing next to one another, is the man standing to the left of the woman, is the woman standing to the right of the man, or are they standing next to one another? 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 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 would change 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. are verbs that could be used. On the other hand, if the event is described from the perspective of the front car, *fleeing*, *leading*, *guiding*, etc. might be used. 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" (Fisher et al. (1994), Lakusta and Landau (2005)). That is, people would 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*.

There are many more examples of visual events that have more than one interpretation, and without the proper context, any of those interpretations could be used to describe that event. 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 (e.g. 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 an early work, Bever (1970) argued that people assumed that the sentences they heard exhibited 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) was the agent and the second NP was 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 verb used to describe an event 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 might not be sufficient for successful performance.

2. Experiment 1

2.1 Methods

2.1.1 Participants

Twenty-one native, monolingual English-speaking undergraduate 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.

2.1.2 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 1920 x 1080 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 Figure 2.1). The horizontal side of the bounding box subtended a visual angle of 36.7° and the vertical side subtended an angle of 20.6°. Four shapes were used (circles, squares, ovals and rectangles) and each scene had two different shapes. The circle subtended a visual angle of 1.15°, and the square subtended a visual angle of 1.00°. The rectangle subtended 1.72° along the longer axis and 0.86° along the smaller one, and the oval subtended 2.12° along the longer axis and 0.72° 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 0.23°/sec. As discussed above, intuitively, what verb is best for describing a visual event depends in part on the animacy of the entities involved (the

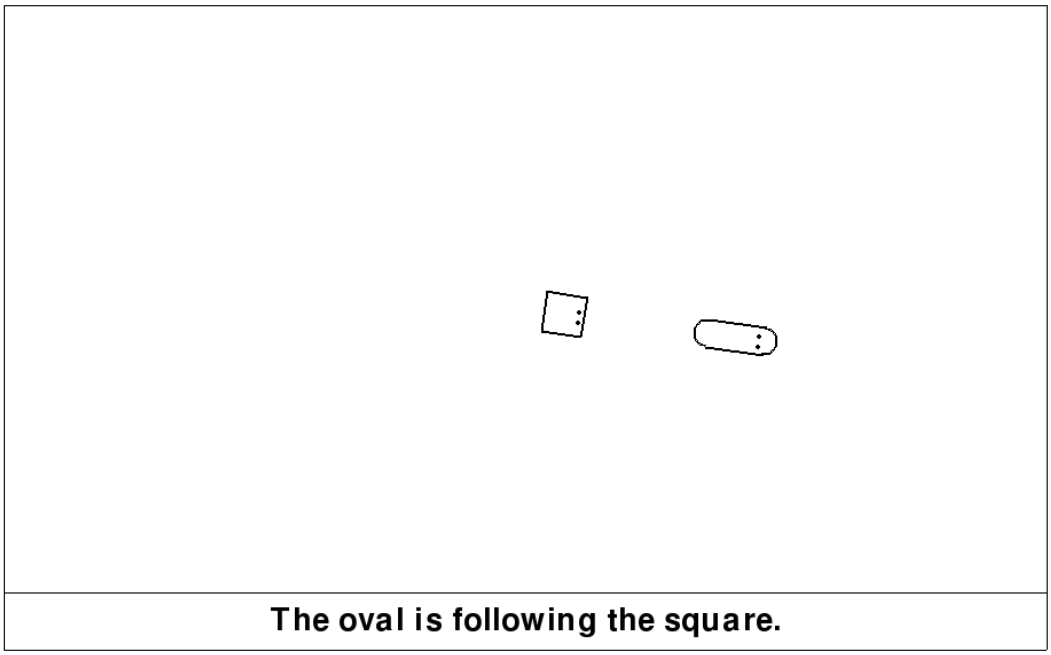

Figure 2.1: A screenshot of the display. (Colors inverted).

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 Figure 2.2, 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 0.29°.

Linguistic Stimuli: As shown in Figure 2.1, 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 with a 40 pt.

Figure 2.2: The No-Eyes (left) and the Eyes (right) conditions.

font size and subtended visual angles between 11.46° and 14.32°. For ongoing actions, the present progressive tense (e.g. *The oval is following the square*) is more semantically 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 inverted. For example, in the trial depicted in Figure 2.1, the match trial caption was, *The oval is following the square*, and the mismatch caption was, *The square is following the oval*.

2.1.3 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.

2.1.4 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 spacebar was pressed until the subject 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.

2.2 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 .50 and .75, it was classified as "weak" evidence; if it was between .75 and .95, it was classified as "positive" evidence; if it was between .95 and .99, it was classified as "strong" evidence; and if it was greater than .99, it was classified as "very strong" evidence.

2.3 Results

Collapsing across all participants' data, participants correctly responded to 5.5% of trials (223 out of 4032). When all trials were included, the mean RT was 3042 ms (SE=38.4 ms), and when only correct trials were included, the mean RT was 2976 ms (SE=35.6 ms), suggesting that there was no speed-accuracy tradeoff.

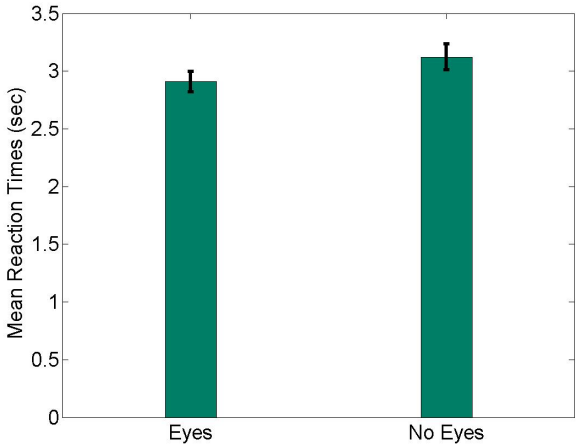

Figure 2.3: Effect of Eyes on RTs. (Error bars = SE).

A 2 (Eyes) × 2 (Match/Mismatch) × 4 (Verbs) ANOVA of correct trial RTs revealed a main effect of Eyes with participants responding about 200 ms (∼7%) faster when the shapes had 'eyes' than when they did not (2909 ms and 3122 ms, respectively; $F(1,20)=8.73$, $p=.008$; Figure 2.3). Bayesian analysis confirmed the result and provided positive evidence favoring the hypothesis that Eyes had an effect ($BF=0.10$, $p_{BIC}(H_1|D)=0.91$).

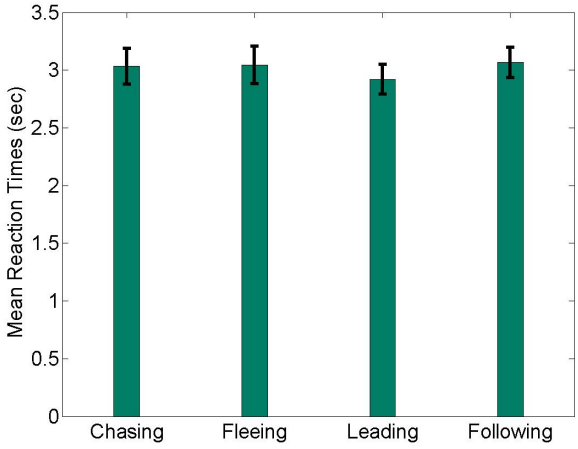

Figure 2.4: Effect of Verb on RTs. (Error bars = SE).

There was no significant effect of verb choice ($F(3,60)=1.00$, $p=0.39$), and Bayesian analysis provided strong evidence confirming this ($BF=57.63$, $p_{BIC}(H_0|D)=0.98$; Figure 2.4). There was no significant difference between the match and the mismatch

conditions ($F(3,60)$=1.00, p=0.39; Figure 2.5), and Bayesian analysis provided positive evidence confirming this (BF=3.56, $p_{BIC}(H_0|D)$=0.78). There were no significant interactions for any independent variables, and Bayesian analyses confirmed these results.

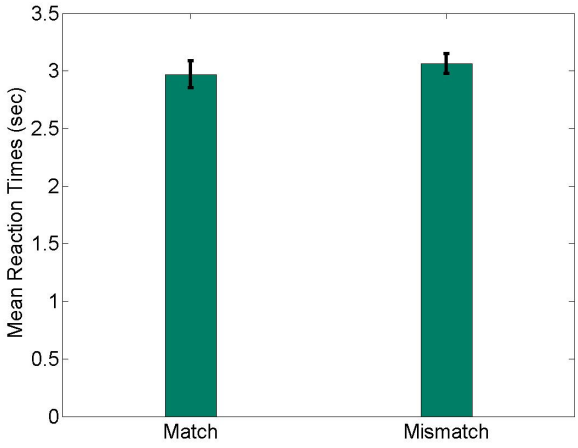

Figure 2.5: Effect of Match/Mismatch on RTs. (Error bars = SE).

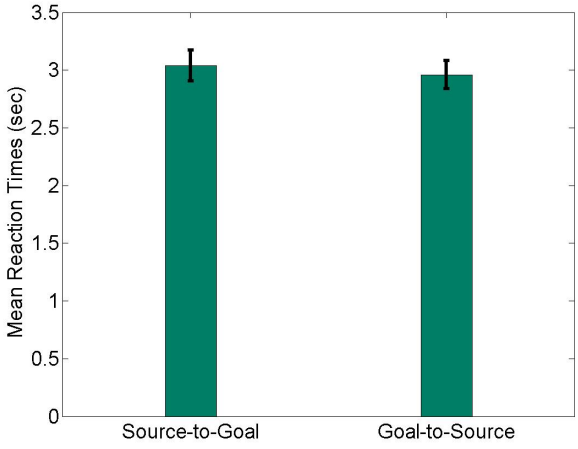

Figure 2.6: Effect of Verb Perspective on RTs. (Error bars = SE).

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 Figure 2.6, a 2 (Eyes) X 2 (Match/Mismatch) X 2 (Verb Perspective) ANOVA of correct trial RTs revealed that Verb Perspective was marginally significant with participants responding about 80 ms

(∼3%) faster for goal-to-source verbs than source-to-goal verbs (2960 ms and 3040 ms, respectively; $F(1,20)=3.62$, $p=0.07$). Bayesian analysis provided only weak evidence in favor of the hypothesis that the two verb groups were different ($BF=0.80$, $p_{BIC}(H_1|D)=0.56$). Verb Perspective did not interact with the other variables, and Bayesian analysis confirmed these results.

2.4 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 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), Tremoulet and Feldman (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 would be processed, animacy obtained, and then the thus-obtained animacy information would be 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 and Eyes argues against animacy being the cause of the Eyes effect.

A second, more plausible, explanation for the Eyes effect is that 'eyes' convey information about the direction of motion in the visual event. In any snapshot of the visual event, the 'eyes' can be used to infer the direction of motion, which, in turn, can be used to determine which shape is in front. This information can subsequently be used to attribute roles to the two shapes. RTs may be greater when 'eyes' are not

present (i.e. the two dots are centric) because the still image is ambiguous and multiple images are required to determine the direction of motion and subsequently process the event.

The lack of an effect of verb or verb perspective could reflect that final linguistic representations are less detailed than the original captions (i.e. they are only "good enough"), perhaps encoding just the details of which entity was where. Alternatively, the lack of a verb effect could be due to the visual and the linguistic systems interacting, and these interactions could have resulted in the semantic contents of the caption influencing the visual representation. If the final visual representation encodes the same information provided by the captions, there would not be a difference between the two representations, and as a result, there would not be an effect of the verbs.

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. However, if that same

template is used to process a different kind of sentence, it might not work. As a case in point, in English, the mapping between grammatical and thematic roles is switched in passive sentences. That is, the subject of a passive sentence is the patient and the object is the agent. Thus, the use of the N V N template for a passive sentence, say, *the square is chased by the circle*, would result in participants incorrectly interpreting the sentence as meaning the square is chasing the circle. Therefore, if participants were using the N V N template as a heuristic to parse the sentences, with the inclusion of passive sentences, they would be incorrect in all trials containing passive sentences.

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. For passive sentences, the model uses lexical information in passivized verbs that signal that the sentence is a passive sentence to correct its initially incorrect parse and to generate a new, revised representation. 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 in case of passive sentences, the system needs to revise its initial incorrect parse, both possibilities predict a difference in 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 half the sentences were actives and half passives.

3. Experiment 2

3.1 Methods

3.1.1 Participants

Twenty native, monolingual English-speaking undergraduate 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.

3.1.2 Stimuli and Apparatus

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

The linguistic component differed from the first experiment in two ways. First, whereas all of the captions in the first experiment were active sentences, in the second experiment, half of the captions were active sentences and half were passive sentences. Superficially, passive sentences like *The oval is chased by the square* differ from simple active sentences like *The oval chases the square* in three ways. First, passives must have a passive auxiliary verb (e.g *is*). [1] Second, in passives, verbs have a passive participle morpheme (e.g., the *-ed* in *chased*). Third, in passives, the preposition, *by*, precedes the object of the sentence (e.g. *by the square*). All of the passive sentences had the form *The* SHAPE_1 *is* VERB*ed by the* SHAPE_2.

As was the case in Experiment 1, in Experiment 2, all of the active sentences were in the present progressive form (i.e. *The* SHAPE_1 *is* VERB*ing the* SHAPE_2). We used the present progressive for two reasons. First, as discussed in Experiment 1, the present progressive tense is the most semantically felicitous tense. Second, using the present progressive in both Experiment 1 and Experiment 2 allows up to compare how actives were processed in the two experiments.

The four verbs used in Experiment 2 were *lead*, *follow*, *guide*, and *trail*. Notice

[1] Note that the '*is*' in a progressive active sentence is different from the passive auxiliary '*is*.' For example, there can be a progressive passive sentence, *the square is being chased by the circle*, where '*is*' is the passive auxiliary and '*being*' is the progressive auxiliary.

that two of the verbs (*lead* and *follow*) were also used Experiment 1 and two were not (*guide* and *trail*). The reason that two were different is that one of the verbs used in Experiment 1 (*flee*) does not passivize (**the square is fled by the oval*). Thus, verbs *flee* and its semantic pair, *chase* had to be replaced. Notice that the replacement verbs *guide* and *trail* can also describe visual events where two entities are moving and one is behind the other. Also, like *chase* and *flee*, *guide* and *trail* are semantic pairs, describing the same event from two different perspectives.

3.1.3 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.

3.1.4 Procedure

The experimental procedure was the same as the first experiment.

3.2 Analysis

Data were analyzed in the same way as they were in the first experiment.

3.3 Results

Collapsing across all participants' data, participants correctly responded to 4.5% of trials (342 out of 7680). When all trials were included, the mean RT was 3557 ms (SE=27 ms), and when only correct trials were included, the mean RT was 3571 ms (SE=27ms), suggesting that there was no speed-accuracy tradeoff.

A 2 (Eyes) X 2 (Match vs. No Match) X 4 (Verbs) X 2 (Syntactic Voice) ANOVA of correct trial RTs failed to reveal a significant effect of Eyes (F(1,19)=1.91,

p=0.18; Figure 3.1), and Bayesian analysis provided weak evidence in support of the hypothesis that there was no effect (BF=1.72, $p_{BIC}(H_0|D)$=0.63).

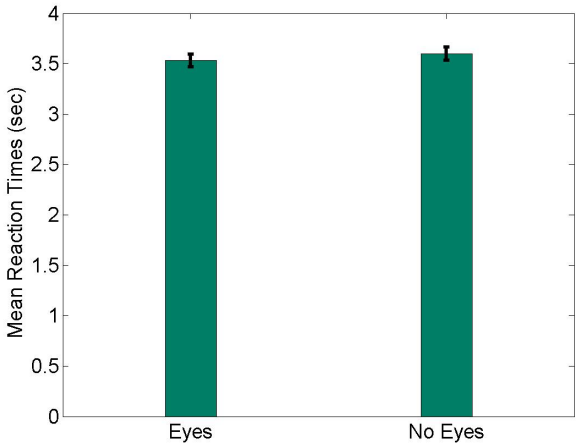

Figure 3.1: Effect of Eyes on RTs. (Error bars = SE).

There was a significant effect of the choice of verb with participants responding fastest when the verb was *lead*, followed by *guide*, and taking more or less the same amount of time for *trail* and *follow* (*lead*=3238 ms, *guide*=3525 ms, *trail*=3729 ms, and *follow*=3764 ms; $F(3,57)$=15.411, $p < 0.001$; Figure 3.2). Bayesian analysis confirmed the result and provided positive evidence favoring the hypothesis that verb choice affected participants' RTs (BF=0.236, $p_{BIC}(H_1|D)$=0.81).

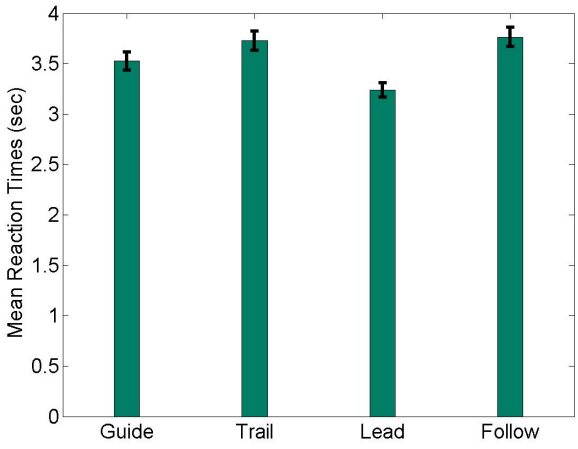

Figure 3.2: Effect of Verb on RTs. (Error bars = SE).

There was also a significant difference between the match and mismatch conditions with participants responding about 280 ms ($\sim$ 8%) faster when the caption matched the video than when it did not (3422 ms and 3706 ms, respectively; $F(1,19)$=26.26, $p < 0.001$; Figure 3.3). Bayesian analysis confirmed the result and provided very strong evidence favoring the hypothesis that the match and mismatch conditions were different (BF=0.0007, $p_{BIC}(H_1|D)$=0.999).

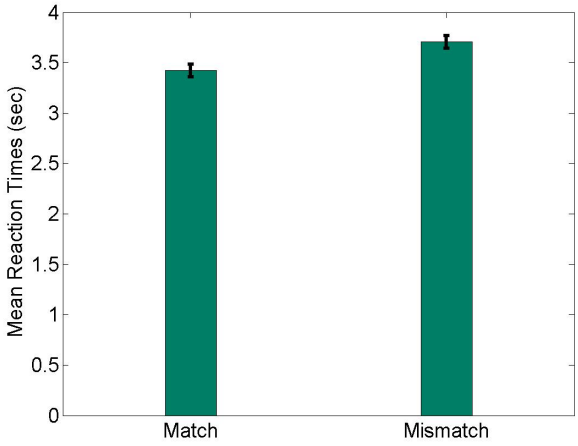

Figure 3.3: Effect of Match/Mismatch on RTs. (Error bars = SE).

Participants were about 480 ms ($\sim$15%) faster on active sentences than passive sentences (3323 ms and 3805 ms, respectively; $F(1,19)$=25.041, $p < 0.001$; see Figure 3.4). Bayesian analysis confirmed this result and provided very strong evidence favoring the hypothesis that the active and passive sentence conditions were different (BF=0.0009, $p_{BIC}(H_1|D)$=0.999).

As depicted in Figure 3.5, there was a significant interaction between verb choice and syntactic voice ($F(3,57)$=3.94, p=0.013). Inspection of Figure 3.5 suggests that the interaction is due to the verb *lead*, and a post-hoc ANOVA revealed that when data from *lead* trials were excluded, the interaction was no longer significant ($F(2,38)$=1.67, p=0.20; BF=7.4, $p_{BIC}(H_0|D)$=0.88). The significance of the interaction between verb and syntactic voice is unclear and Bayesian analysis provided weak support for the hypothesis that there was no interaction (BF=1.628, $p_{BIC}(H_0|D)$=0.62).

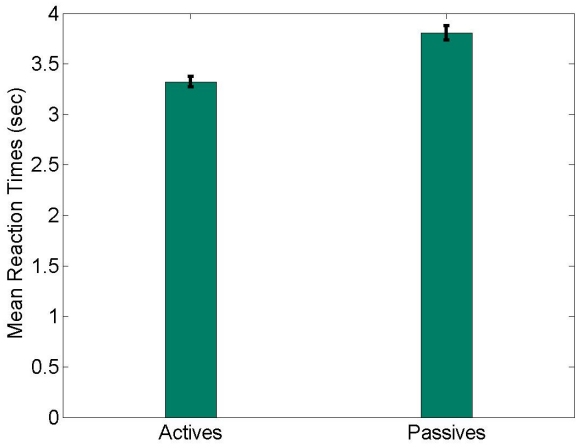

Figure 3.4: Effect of Syntactic Voice on RTs. (Error bars = SE).

There was also a significant interaction between verb choice and match conditions ($F(3,57)$=4.82, p=0.005; Figure 3.6) and Bayesian analysis provided weak evidence supporting the interaction (BF=0.525, $p_{BIC}(H_1|D)$=0.66). Inspection of Figure 3.6 suggests that the interaction is possibly 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 ($F(2,38)$=1.17, p=0.32; BF=12.07, $p_{BIC}(H_0|D)$=0.92).

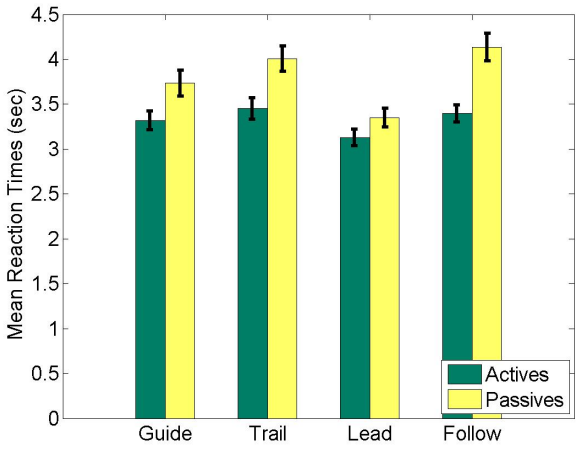

Figure 3.5: Interaction between Verb and Syntactic Voice. (Error bars = SE).

There were no other significant interactions, and Bayesian analyses confirmed these results.

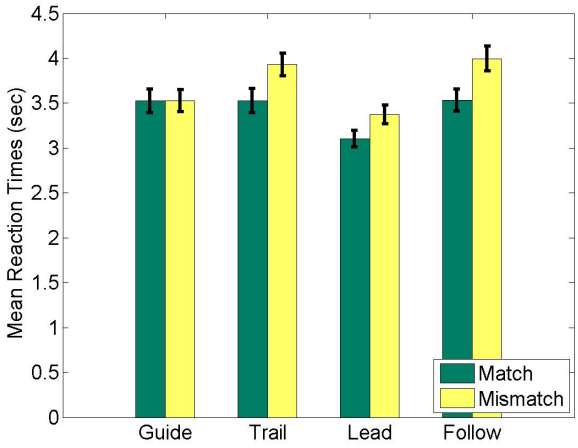

Figure 3.6: Interaction between Verb and Match/Mismatch. (Error bars = SE).

As in Experiment 1, the verbs were grouped as "source-to-goal" verbs (*trail* and *follow*) or "goal-to-source" (*guide* and *lead*). A 2 (Eyes) X 2 (Match vs. No Match) X 2 (Verb Perspective) X 2 (Syntactic Voice) ANOVA of correct trial RTs revealed a significant effect of Verb Perspective with participants responding about 360 ms (∼11%) faster when the verbs belonged to the goal-to-source group than when they belonged to the source-to-goal group (3382 ms and 3746 ms, respectively; $F(1,19)=24.098$, $p < 0.001$; Figure 3.7). Bayesian analysis confirmed the result and provided very strong evidence in favor of the effect ($BF=0.001$, $p_{BIC}(H_1|D)=0.999$).

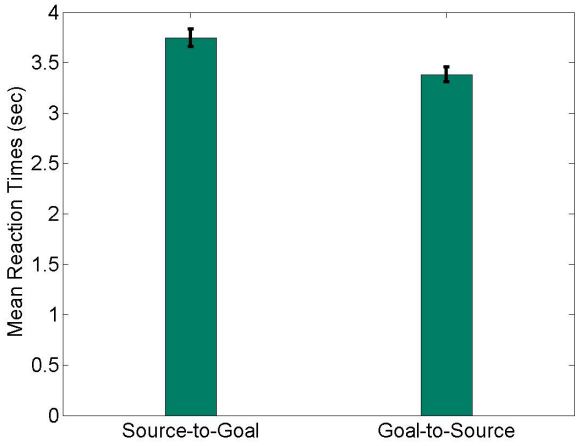

Figure 3.7: Effect of Verb Perspective on RTs. (Error bars = SE).

As shown in Figure 3.8, there was a significant interaction between Verb Perspective and syntactic voice ($F(1,19)$=7.263, p=0.014), and a Bayesian evaluation confirmed the result and provided positive evidence supporting the interaction (BF=0.176, $p_{BIC}(H_1|D)$=0.85). Inspection of Figure 3.8 suggests that the interaction is the result of a greater "cost" for passive sentences for source-to-goal verbs. Segregating the data by Verb Perspective revealed that there was a difference in the level of the effect of syntactic voice for the two verb perspective groups. The difference between the mean RTs corresponding to active and passive sentences was 651 ms for the source-to-goal group and 320 ms for the goal-to-source group. However, both differences were significant ($F(1,19)$=26.485, $p < 0.001$; $F(1,19)$=10.563, p=0.004, respectively).

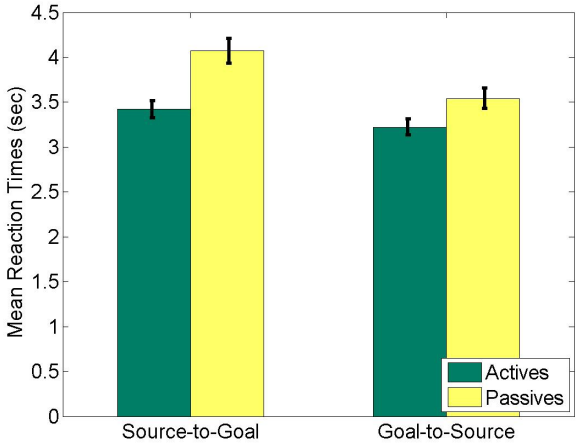

Figure 3.8: Interaction between Verb Perspective and Syntactic Voice. (Error bars = SE).

There was also a significant interaction between Verb Perspective and the match conditions ($F(1,19)$=14.12, p=0.001; Figure 3.9). Bayesian analysis confirmed the result and provided strong evidence supporting the interaction (BF=0.02, $p_{BIC}(H_1|D)$=0.98). Inspection of Figure 3.9 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 434 ms for source-to-goal verbs and 133 ms for goal-to-source verbs. Segregating the data by verb perspective revealed that the difference between the match and mismatch conditions was significant only for the source-to-goal verb class ($F(1,19)$=41.858, $p < 0.001$;$F(1,19)$=3.842, p=0.065, respectively).

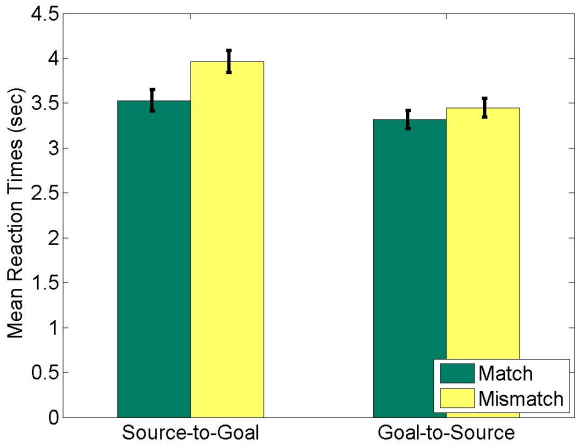

Figure 3.9: Interaction between Verb Perspective and Match/Mismatch. (Error bars = SE).

There were no other significant interactions, and Bayesian analyses confirmed these results.

3.4 Discussion

To a first approximation, the results of the second experiment are complementary to the results of the first: In the first experiment, the presence or absence of 'eyes' was the only factor that had an effect, and in the second experiment, it was the only factor that did not have a significant main effect. That suggests one of two possibilities. Eyes may not have had an effect on the participants in Experiment 2, but given that the visual stimuli were the same as in Experiment 1, this seems unlikely. Alternatively, Eyes did have an effect in Experiment 2, but the effect was too transient to be detected in the end-of-trial RTs.

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 syntactically 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, these differences between the verbs may not have been encoded. As a result, comparisons between the visual and the linguistic descriptions would not reflect inter-verb differences.

Previous work (e.g. Fisher et al. (1994), Lakusta and Landau (2005)) suggests that people may 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 say that, in our study, goal-to-source verbs are better descriptors for the events depicted in our videos. One of the subtleties of our visual display was that near the corners, the shape that was behind would slow down and allow the shape in front to move away. This behavior was implemented to ensure that the shape that was behind never "catches up" with the shape in front. It may be that this behavior made it seem as if the shape that was behind let the shape in front "guide" it or "lead" it as they moved along.

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 indiate that when adult participants are asked 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 of them. 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

very accurate. It may be that other factors such as the frequency of occurrence of linguistic items, influence the way we talk about events, even if there are other, better ways of describing an event, using less frequent words. As a case in point, the source-to-goal verbs used in our study (*chase*, *follow*, *trail*) occur more frequently in the Kucera and Francis (1967) database than the goal-to-source verbs (*flee*, *lead*, *guide*).

Furthermore, the fact that verbs interacted with syntactic voice and match/mismatch conditions also suggest that the four verbs are processed differently. Our finding that the cost associated with processing passive sentences was not as high for goal-to-source verbs as for source-to-goal verbs suggests that for the former verbs, the process of revision was quicker. Perhaps, source-to-goal verbs form stronger initial representations, possibly because they occur more frequently than goal-to-source representations. Our finding that the cost of verification in case of a mismatch 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.

4. General Discussion

Taken together, the results of the two experiments presented here suggest that people form different types of linguistic representations depending on the task that they are asked to perform and the detail necessary to perform it. In Experiment 1, participants only needed a simple heuristic parser to process the sentences used as captions. Our conjecture is 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. In Experiment 2, a purely heuristic parser does not provide a detailed enough representation to correctly interpret passives, and, hence, a more detailed, syntactic analysis is required to accurately perform the task. Our conjecture is 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.

Another interesting post-hoc revelation is that participants on average responded about 350 ms faster in Experiment 1 than in the actives-only trials of Experiment 2. That is, participants were slower for syntactically identical sentences in Experiment 2 than Experiment 1. This further suggests that sentences were processed differently in the second experiment.

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 the used 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 does the comprehension system decide which of the two (or more) parses of the input is to be used for the 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. The system could start with a "default" behavior: it could either always use the output of the heuristic parser or always the algorithmic. If the default is the heuristic parser, the system will always make an error when a sentence does not conform to the canonical template used. 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.

Bibliography

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. and 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*, pages 279–362. New York: Wiley & Sons, Inc.

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

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

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

Dittrich, W. and 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., and Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11:11–15.

Ferreira, F. and 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., and 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., and 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., and Kaufman, L. (1995). *Distinguishing between animates and inanimates: Not by motion alone*, pages 150–184. Oxford, England: Clarendon Press.

Knoeferle, P. and 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*, pages 1166–1171, Stresa, Italy.

Knoeferle, P., Crocker, M., Scheepers, C., and 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. and Francis, W. N. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

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

MacDonald, M. C., Pearlmutter, N. J., and 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. (in press).

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

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

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

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

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

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

Trueswell, J., Tanenhaus, M., and 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., and 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.