

# **Spatial Language and Temporal Cognition\***

**David January**

**Swarthmore College**

**December 10, 2002**

## **1. INTRODUCTION**

Does the language we speak affect how we think? From investigating whether linguistic color categories determine how those colors are perceived to exploring how a language's system of spatial description influences a person's spatial memory, this question has a long and controversial history. The assertion that language does indeed affect how we think is known as the Linguistic Relativity Hypothesis (also known as the Sapir-Whorf Hypothesis after the scholars who developed it early in the Twentieth Century). In its strong form, the Linguistic Relativity Hypothesis holds that language determines what we are capable of thinking about and how we think about it. The more moderate form asserts that language influences the ways in which we are accustomed to thinking about concepts and ideas, though it does not prevent an individual from thinking in a different way, as the strong form implies. This thesis aims to provide a brief history of the research into the Linguistic Relativity Hypothesis and to provide some new data pertinent to the question of whether the language we speak affects how we think. Specifically, I will report the results from a number of experiments that investigate whether—and how—the fairly universal linguistic treatment of time as a spatial concept shapes the

---

\* I would like to thank the members of my thesis discussion class, Ted Fernald, Steve Wang, and especially Edward Kako for their commentary and insights on this paper. I would also like to thank the Howard Hughes Medical Institute for its financial support of the research for this project.

mental conception of time. The experiments reported below examine if, because we talk about temporal relations as if they were spatial, we think of time as a spatial concept.

## 1.1 HISTORY

### 1.1.1 Whorf and Hopi

Most commonly associated with the works of Benjamin Lee Whorf (Whorf, 1939, 1941, 1956a, 1956b), the Linguistic Relativity Hypothesis has been addressed in many ways, with a variety of results. Whorf is most famous for his work with Hopi, the language of a southwestern Native American tribe of the same name. Whorf's work with the Hopi language led him to believe that the Hopi did not and could not conceive of time in the way that English-speaking people do, and vice versa, because the Hopi language does not discuss time in a way even remotely related to the way English does. Whorf's argument went like this: In English, we talk about time as divided into the past, present, and future in our tense system. Hopi, he continued, did not have this same grammatical distinction among the past, present, and future, instead relying on aspects (predominately the perfective and imperfective; See Jeanne (1978, p. 163), voices, and other morphological markers to convey, for example, whether the speaker knows a situation to be (corresponding to the English past or present) or if he or she expects it to arise (corresponding to the English future). The sentences in Table 1 serve to illustrate Whorf's point: the two tenses in the English sentences are expressed by aspect markers in

Hopi. These differences in the grammatical handling of time result in strikingly different conceptions of

Table 1: English and Hopi sentences. Adapted from Jeanne (1978:170-171)	
ENGLISH	HOPI
the man ran THE MAN RUN+PAST	mi' taaqa wari THAT MAN RUN+PERF
the man is running THE MAN BE+PRES run+PROG	mi' taaqa wari-ki-wta THAT MAN RUN+ki+IMPERF

The ' in the Hopi sentences indicates the glottal stop.

time, he claimed.

English speakers, because of the tendency to talk of time and durations of time in a manner analogous to the treatment of mass nouns<sup>1</sup>, think of time as divided into three chunks corresponding to the grammatical separation into past, present, and future. Speakers of Hopi, by contrast organize time in terms of events getting later and later without such a division, though each type of event “gets later” in a way that is appropriate for the participants in the event. As Whorf says:

Events are considered the expression of invisible intensity factors, on which depend their stability and persistence, or their fugitiveness and proclivities. It implies that existents do not ‘become later and later’ all in the same way; but some do so by growing like plants, some by diffusing and vanishing, some by a procession of metamorphoses, some by enduring in one shape till affected by violent forces. (Whorf, 1939, p. 147).

As evidence for his claim, Whorf offered a long anthropological analysis of both Hopi and European culture and claimed that the differences in them are a necessary result of the different ways of thinking about time that arise from the different ways of talking about time. For example, Whorf claims the Hopi preoccupation with “preparing”, of objects, friendships, etc., results from the more cyclic treatment of time grammatically. The Western propensity to look at the universe as dualistic<sup>2</sup>, as composed of substance and container; the tendency to look at the soul as a “substance” and the body as a “container” comes from the grammatical treatment of mass nouns.

---

<sup>1</sup> For example, *milk*, a mass noun, can be referred to as *a carton of milk*. *Time* also follows this pattern: *a period of time*, that will take *a lot of time*.

<sup>2</sup> Consult Whorf (1939) for further discussion.

Whorf's work with Hopi has a number of problems, however. First, his descriptions of Hopi structures are very vague, and he does not formulate concrete, testable predictions for the Linguistic Relativity Hypothesis. A second problem with his work is that it is impossible to determine whether the differences in language result in differences in culture or, conversely, differences in culture result in differences in language. The Hopi might speak of time more cyclically than Westerners because their culture treats it as a cyclical thing. Further, more recent linguistic analysis of Hopi and English has shown that most of Whorf's analysis was wrong. For example, English only has 2 tenses: the present and the past (e.g., *he jumps*, *he jumped*). The future in English is expressed through the use of modal auxiliary verbs (e.g., *he will jump*), and Whorf's claim that "the Hopi language gets along perfectly without tenses for its verbs" (Whorf, 1956b) seems to be incorrect. Hopi does have two tenses: the future and non-future (for further discussion, see Jeanne, 1995). A more serious problem for the theory, however, is that Whorf does not offer any psychological proof of his claims. His descriptions of differences in thought patterns are all based in the linguistic patterns he observes, which is the very thing he claims is the cause of the differences in thought.<sup>3</sup>

### 1.1.2 Color Terms

Though Whorf's initial work with Hopi has long been viewed as inadequate proof of his Linguistic Relativity Hypothesis, the question of whether native language has an effect on thought has received a great deal of research attention. Much of the early work in the area grew out of the cross-linguistic differences in color terms. Different languages make use of different basic color terms to divide up the wavelengths of visible light. Berlin and Kay (1969) found that languages can have as few as two basic color terms, or as many as

ten or eleven. For the purposes of their study, “basic color term” was defined as a monolexemic word, the use of which was not restricted to a limited class of items, and that is not considered a subset of some other color term (e.g., *crimson* is part of the *red* colors, and therefore does not count as a basic color term). Further, Berlin and Kay required that the term be “psychologically salient” for their informants (Berlin and Kay, 1969, p. 6). This means that it was used with a great deal of consistency across informants and that each informant used the term<sup>4</sup>. Using these criteria, English has 11 basic color terms: *black*, *white*, *red*, *green*, *yellow*, *blue*, *brown*, *purple*, *pink*, *orange*, and *gray*. These terms are predominately based on differences in hue among the color categories. By contrast, the Dani of Indonesia have two color terms that are predominately based on differences in brightness between the color categories: *mili*, referring to the dark, cold colors, such as blue, and *mola*, referring to the bright, warm colors, such as pink.

Recognizing the fact that color terminology differs across languages, researchers thought that they had found an area rich in support for Whorf’s hypothesis: the different linguistic ways of dividing the color spectrum should, if the Linguistic Relativity Hypothesis is correct, result in different psychological experiences of the color spectrum. This would also be a very informative area for research because it was so far removed from Whorf’s style of “proof” through cultural analysis. The sensation of the color spectrum is not easily affected by culture; no matter what culture people live in, they should have similar structures in the eye and brain for perceiving color. Any differences in perception would, therefore, have to be the result of the processing of the information

---

<sup>4</sup> Other criteria were used to make finer-grain judgments of what counted as a basic color term. See Berlin and Kay (1969) for a full list.

presented to the rods and cones in the eye. This effect on processing would be proof of the effects of Linguistic Relativity.

At first, the results seemed promising. Brown and Lenneberg (1954) found that codability of colors was positively correlated with accuracy of recognition, where “codability” was defined as a composite of agreement in naming, length of name, and response latency of naming. The codability of any give referent could change from language to language; an object with a one-word name in English might need an entire phrase to be properly identified in another language, or vice versa (e.g., the German *schaudenfreude* translates, in English, to *joy at another’s suffering*). Since codability is a linguistic measure and recognition accuracy a psychological one, it seemed that Brown and Lenneberg had finally offered some psychological proof in support of the Linguistic Relativity Hypothesis: things more easily coded linguistically are more easily remembered. However, Heider and Olivier (1972) performed a study that has come to be respected as the classic refutation of the Linguistic Relativity Hypothesis.

Heider and Olivier (1972) took speakers of English, with their 11 color terms based on hue, and speakers of Dani, with their 2 color terms based on brightness, and compared them in a color memory task. Subjects were shown an initial color chip, and then 30 seconds later they were shown a test array with the chip they had seen before and a set of distracters they had not<sup>5</sup>. The Linguistic Relativity Hypothesis would predict that, because the Dani color naming system is based on brightness and the English color naming system on hue, Dani and Americans would differ as to how accurately they could make discriminations along the different color dimensions.

Heider and Olivier found this not to be the case. Their results show that, within each group, the Dani and the Americans did not vary as to how often group members confused a perceptually-adjacent distracter tile with the tile they initially saw when the tiles differed only in their hue or only in their brightness. This result held even when the hue differences between the two tiles caused a name difference for the English speakers and when the brightness differences caused a name difference for the Dani. Further, Heider and Olivier found that the apparent structure of the naming space for color, different in each language, did not result in a different structure of the memory space for the different languages. To construct a model of the naming space for color, they had subjects name color chips in a large array. The memory space was modeled through the number of times a perceptually-adjacent chip was confused for a target chip. Assuming that if relativistic effects did not evidence themselves in color, they would not evidence themselves in any domain, Heider and Olivier concluded that there is no effect of language on thought.

### 1.1.3 Space

Because color perception is so strongly grounded in the physical responses of the rods and cones in the eye, researchers now believe that it is not the best domain to examine for Whorfian effects. The physical reactions of the cells in the eyes and the structures in the brain, they argue, would be particularly resistant to linguistic influences. For this reason, research in recent years has turned to other domains to search for evidence of Linguistic Relativity.

---

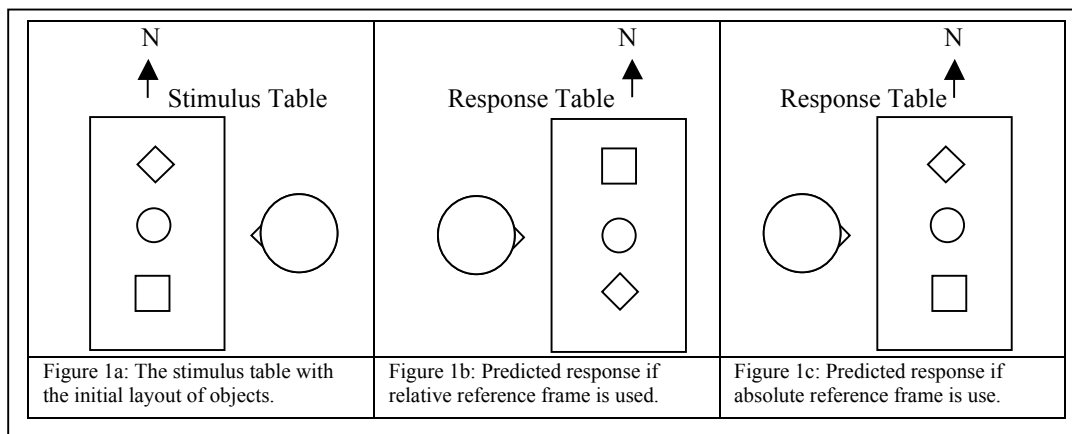
<sup>5</sup> The chips were Munsell chips, which are specially constructed to vary in color, brightness, or hue by equal gradations to the observer.

The domain of space and spatial reasoning is one area that has received a great deal of research attention. Space is a particularly good domain for investigation into the Linguistic Relativity Hypothesis because there are many different ways cross-linguistically to discuss space and spatial relationships, just as there are many ways to divide up the color spectrum. Space is a better candidate than color for relativistic effects, however, because spatial understanding is based largely, if not wholly, on the mental representations of spatial relationships. Where color sensation was grounded in physiology, perception of spatial relationships is almost entirely a mental phenomenon, and as a result one can expect spatial representations to be highly susceptible to the effects of Linguistic Relativity, or at least *more* susceptible than color sensation.

Stephen C. Levinson and his colleagues have noticed that there appear to be three distinct ways of talking about space—three spatial reference frames—that are used to locate a target object with respect to a background object: the intrinsic reference frame, the relative reference frame, and the absolute reference frame. The intrinsic reference frame uses a designated facet of the ground object to locate the target object. For example, “the ball is *in front of* the chair” locates the ball with respect to the chair’s own front. The relative reference frame locates the target object in relation to the background object from the speaker’s perspective, e.g., “the ball is *to the right of* the chair” means that the ball is on the speaker’s right. Finally, an example of the absolute reference frame would be the cardinal directions. “The ball is *north of* the chair” is true, no matter how the chair is facing or how the speaker is viewing the relationship between the ball and the chair (Levinson 1996, 2001). The following is an example of the debate over Linguistic Relativity in the domain of space.



Levinson (2001) took speakers of Dutch, in which the relative and intrinsic reference frames are most often used<sup>6</sup>, and speakers of Tzeltal, in which the absolute system is the only system in use, and compared how they performed on a spatial reasoning task. Given three toy animals in a row on a table, each participant was rotated 180° and asked to reproduce the layout of animals on a second table. Upon seeing the table, Dutch speakers would most probably say something that translates to “the duck is *on my right*”, whereas the Tzeltal speakers would most likely say something that translates to “the duck is *north of me*”. The rotation in the experiment is important because use of the different reference frames (relative and absolute) results in different interpretations of what is “the same”: Dutch speakers should recreate the array of animals with the duck on their right, even though this now makes the duck south of them, whereas Tzeltal speakers should recreate the array of animals with the duck north of them, even though this makes the duck on their left. Figures 1a-c provide a more concrete example of these predicted differences. If the relative reference frame is used, then if given a



layout as in Figure 1a, one should reproduce it as in Figure 1b; if the absolute reference frame is used, then one should reproduce Figure 1a as in Figure 1c.

<sup>6</sup> Dutch does use all three reference frames, but the absolute system does not come up in common parlance very often and tends to be used only in navigational contexts.

Levinson (2001) found that Dutch speakers responded consistent with the relative reference frame and Tzeltal speakers responded consistent with the absolute reference frame. Since Levinson believed the only difference between these two populations of significance was that Dutch speakers use a relative reference frame in their language and Tzeltal speakers use an absolute reference frame, he concluded that these differences are a result of the differences between the languages.

Li & Gleitman (2002), however, believed that environmental differences at the time of testing might have influenced responses. They argued that the fact that Tzeltal speakers were tested outside, next to a large building and Dutch speakers inside with the shades down makes comparisons between the two populations invalid. To test this, they took only native speakers of English, who also use a relative reference frame, and varied the environment in which they took Levinson's test. Li & Gleitman found that they could make English speakers respond like Tzeltal speakers if they varied the environment under which they were tested. If English speakers were tested inside, with the shades down, as the Dutch were in Levinson's experiment, they responded as if they used the relative reference frame. However, if English speakers were tested outside next to a large high-rise apartment building they were much more likely to respond as if they were using the absolute reference frame. Thus, Li & Gleitman conclude that it was not the language that caused the differences in response, since all subjects were English speakers, but rather the differences in environment.

Levinson *et al.* (2002) counter this line of argument, providing some additional experimental evidence for their claim. First, they attempted to replicate Li & Gleitman's (2002) results with Dutch speakers at the University of Nijmegen, but were unable to, for

reasons they cannot explain. They contend, however, that the fact that the Li & Gleitman study did not have any additional load (beyond the rotation) might be partly responsible. In the original Levinson (2001) study, subjects had to select the toy animals that were on the stimulus table from a set of four animals for the response table, which “reduces the chance of participants second-guessing the purpose of the experiment, and increases the chance of participants falling back on their habitual default frame of reference” (Levinson *et al.*, 2002, p. 164). They next ran an experiment in which a toy man was moved along a maze path, and subjects were asked to replicate that path on a response table under the same 180° rotation as in the “animals-in-a-row” task, with both an absolute and relative response available. Levinson *et al.* compared Dutch-speaking subjects both indoors and outside with large buildings nearby—an environment similar to the outside environment in Li & Gleitman (2002). The comparison between the inside and outside conditions revealed no significant difference, giving greater support to the idea that the role of load was important in the animals-in-a-row task.

Levinson *et al.* (2002) further claim that Li and Gleitman (2002) misunderstood the test between the two reference frames and for this reason came to an erroneous conclusion. The three reference frames for discussing spatial relationships (discussed above) not only make different predictions about what is “the same” when the speaker rotates, they also make different predictions when the ground object—the object that the target object is located in reference to—rotates. This fact, they argue, is necessary to disambiguate between what Levinson *et al.* (2002) call the allocentric reference frames: intrinsic and relative. When the subject is rotated 180°, the two allocentric reference frames do not necessarily make different predictions about what’s the same.

One variation on the animals-in-a-row experiment run by Li & Gleitman (2002) involved the inclusion of a small biasing landmark (a plastic duck pond) on the stimulus and response tables. Li & Gleitman placed these landmarks either always on the subjects' right (the relative condition) or on the south end of both the stimulus and response tables (the absolute condition), resulting in a left/right alternation under rotation. Under this manipulation, Li & Gleitman found that subjects responded in a relative manner when the duck ponds were placed relatively and absolutely when the duck ponds were placed absolutely, giving, they argue, greater strength to their claim that the local environment is crucial in determining how to encode the spatial relationships among objects. Levinson *et al.* (2002) claim that duck ponds were not treated by subjects as landmarks, but rather as part of the scene to replicate because they are obviously movable. Thus, they claim, subjects in Li & Gleitman's experiment were not making use of the absolute coordinate system when they faced the animals to the south, but rather were using the intrinsic reference frame, with the duck pond serving as the ground object the other objects were arranged in relationship to.

To support this claim experimentally, Levinson *et al.* (2002) ran the same duck pond experiment under 90° rotation. This quarter rotation, they claim, will fully disambiguate between the two reference frames (intrinsic and absolute) because, if subjects were responding absolutely, then they would arrange the animals with the duck pond as the south-most member. If, however, they are responding intrinsically, subjects would place the animals in an East-West line with the duck pond occupying the same spot in the array on the stimulus and response tables (See Figure 2). Dutch speakers, who make use of the same reference frames as English speakers, in this experiment responded in accordance

with the intrinsic reference frame. This result is taken to show that Li & Gleitman (2002)

failed to disambiguate between the intrinsic (objects organized around another object)

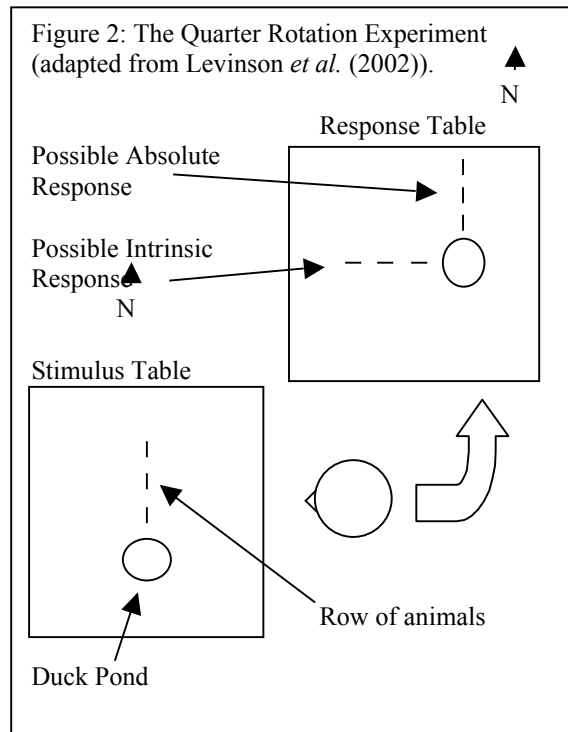
and absolute (objects organized with respect to a coordinate system) reference frames in their experiment, with the result that they misinterpreted their results.

Instead of English speakers being influenced by their environment to switch between the relative reference frame, commonly in use, and the absolute reference frame, in use only in special

contexts, Levinson *et al.* (2002) claim that English speakers were alternating between the intrinsic and relative reference frames, both commonly in use in English.

## 1.2 TIME AS A TEST OF THE WHORFIAN HYPOTHESIS

As the previous example of investigation of Whorfian effects suggests, the Linguistic Relativity Hypothesis is still hotly debated, even if only in its weaker forms. The widespread use of metaphors in language provides another opportunity for research into the Linguistic Relativity Hypothesis. Lakoff and Johnson (1980) found that metaphors<sup>7</sup> form an integral part of language. Moreover, they found that these metaphors often come as a structured, systematized unit, not as random comparisons between two



domains. For example, there are numerous exemplars of the metaphor LOVE IS A JOURNEY: “look *how far we’ve come*”, “it’s been a *long, bumpy road*”, “we’ve gotten *off the track*”, “our marriage is *on the rocks*”. All of these various exemplars, however, fall under one heading, namely, LOVE IS A JOURNEY, which makes them a coherent system. (Lakoff and Johnson, 1980).

Another area in which metaphors of this type are common is in the linguistic treatment of temporal relationships. Languages often make use of a spatial metaphor when describing temporal relationships. In English we see this in such sentences as “Tuesday comes *before*<sup>8</sup> Wednesday”, “she left *after* he did”, and “The hard times are *behind* us”, “I’m looking *forward* to a brighter tomorrow”.

This area is particularly fruitful for research into the Linguistic Relativity Hypothesis because people only have limited direct experience of time. People learn from experience that they can only be in one place at one time, that each moment only happens once, and that they cannot go back to an earlier point in time. This common experience has led to a consistent cross-cultural conception of time as a “one-dimensional, directional entity” (Boroditsky, 2001). However, though there does seem to be some universal experience of time, there are many elements of the concept “time” that are not directly available to experience. Given that time is a directional entity—a concept with motion inherently a part of it—it can either move past people or people can move through it. Given that time is unidimensional, it can run either horizontally or vertically

---

<sup>7</sup> In this sense, metaphor does not refer to figurative language but rather the between-domain comparisons of everyday speech. Ex: “Things are looking up” relies on the metaphor GOOD IS UP. See Lakoff and Johnson (1980) for more examples.

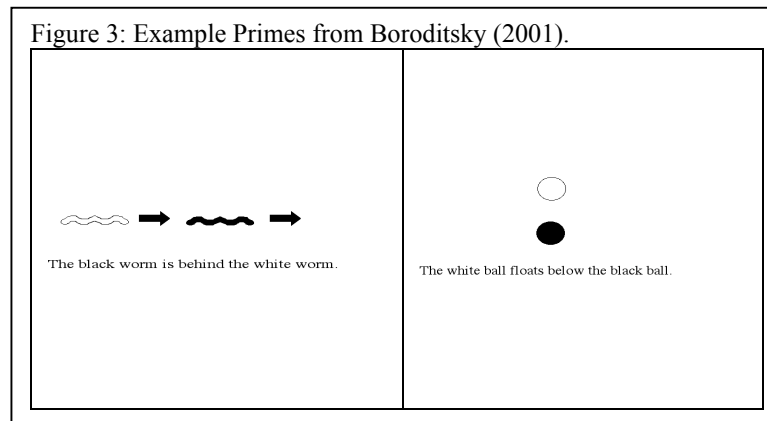
<sup>8</sup> Some readers might object that *before* and *after* are solely temporal terms. This objection is addressed below.

in space. Individual languages can vary freely on these aspects of the concept “time” when talking about time and still satisfy the universals of temporal experience.

(Boroditsky, 2001).

Because there can be such variation among languages in how they treat time, speakers of those languages could also have different conceptions of time—if the Linguistic Relativity Hypothesis is correct. Indeed, Boroditsky (2001) performed a study in which she compared native speakers of English, who talk of time horizontally (e.g., “Tuesday comes *before* Wednesday”), and native speakers of Mandarin, who talk of time vertically (e.g., “Tuesday comes *above* Wednesday”), in a task wherein subjects were

shown a series of slides and asked to indicate whether the sentences on the slides were true or false. Some of the slides were of either



horizontally- or vertically-oriented spatial relationships, which were intended to prime horizontal reasoning or vertical reasoning (see Figure 3 for examples). Next, subjects were shown a statement about the ordering of two months in a year and asked to indicate whether the statement was true or false. Reaction times to the temporal ordering questions were measured and compared between populations and across the two types of prime slides.

Boroditsky found that English speakers could answer these temporal questions faster after horizontal primes than after vertical primes and that for Mandarin speakers

the pattern was reversed, being faster after vertical primes than horizontal primes. Since the only relevant difference between these two populations is that English speakers talk of time horizontally and Mandarin speakers talk of vertically, it seems that it must be that the differences in the linguistic treatment of time resulted in differential priming effects for temporal reasoning. Further, since there was this vast difference in response pattern to targets that used the *before/after* opposition and those that used the *earlier than/later than* opposition for native Mandarin speakers, it seems that there must be some conceptual difference between these two sets of terms.

Boroditsky's Mandarin speakers were actually Mandarin-English bilinguals (all with Mandarin as their native language). She chose to test a bilingual population in its non-native language in order to test for effects of the first language on processing for tasks independent of that language, like processing in a second language. Since the Mandarin-English bilinguals still responded reliably differently from native English speakers, even when tested in English, Boroditsky concludes that there is a lasting effect of native language on the conception of time that is formed early on by the first language and that second languages do not erase.

In a third experiment, Boroditsky "trained" English speakers to talk about time making use of the vertical metaphor (e.g., "Monday comes *above* Tuesday"). In this third experiment she found that trained English speakers responded faster to targets after *vertical* primes, reversing the untrained-English-speaker pattern, and claims that the new way of talking about time has affected how English speakers conceive of it. This is in contradiction to her findings in the second experiment, where she found that the introduction of a new way of talking about time does *not* affect the conception of time.



The above contradiction in Boroditsky's study casts some doubt on the strength of the effect she found. Coupled with the fact that Boroditsky's relativistic claims contradict earlier findings in the literature (e.g., Heider and Olivier (1972), discussed above), there is a need to repeat Boroditsky's experiment to test the robustness of the effect she found, which we do below with English speakers.

## **2. THE EXPERIMENTS**

### **2.1a Experiment 1a:**

#### **Methods:**

Methods in Experiment 1a were modeled after Boroditsky (2001), except all participants were speakers of English.

#### *Participants:*

Participants were 24 students at Swarthmore College. All subjects had learned English by age 5 and can be treated as native speakers. Subjects were offered \$5 and candy for participation in the experiment.

#### *Design:*

Subjects completed 32 experimental trials and 32 filler trials composed of two spatial prime questions and then a target question about a temporal relationship. Primes were spatial scenarios accompanied by a sentence description as described below. Target questions presented a temporal relationship using either spatial (before/after) or temporal (earlier/later) comparative terms. The experimental trials were arranged such that the first prime question was false, the second prime question true, and the target question true, to follow Boroditsky (2001). Filler trials had the same structure as the experimental trials, except the target temporal question was always false, and the true/false order of the

primes was randomly determined. Experimental trials and filler trials were randomly interspersed to prevent subjects from discovering this pattern for experimental trials. Participants answered each target question twice, once after vertical primes and

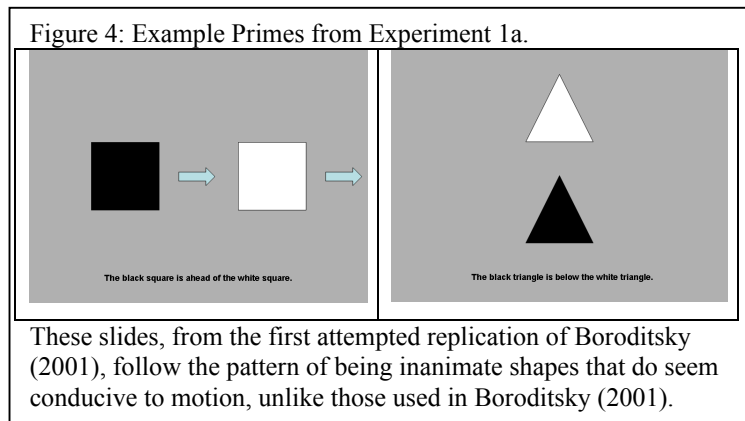
<b>Prime Type:</b> Horizontal	<b>Prime Type:</b> Horizontal
<b>Target Type:</b> Spatial	<b>Target Type:</b> Temporal
<b>Prime Type:</b> Vertical	<b>Prime Type:</b> Vertical
<b>Target Type:</b> Spatial	<b>Target Type:</b> Temporal

once after horizontal primes. The order of trials was randomized newly for each subject. The experiment had a fully crossed within-subject 2 (prime type) x 2 (target type) design.

*Materials:*

A set of 128 primes and 16 targets and 16 fillers, all TRUE/FALSE questions, was constructed.

*Primes:* The 128 primes were composed of horizontally- and vertically-oriented primes. Half of the horizontal primes used the “X is before Y” construction,



the other half used the “X is behind Y” construction, with the left/right orientation balanced. Half of the vertical primes used the “X is above Y” construction; the other half used the “X is below Y” construction. Autoshapes from Microsoft PowerPoint were used to create primes (see Figure 4 for examples). Half of each group of primes were arranged to elicit a T (true) response, half an F (false) response.

*Targets:* 16 statements about the order of the months were constructed. Half of these use the spatial comparative terms *before* and *after* (as in “June comes *before*

August”), half used the purely temporal comparatives *earlier* and *later* (as in “March comes *later* than February”). All of these target statements were true.

*Fillers:* 16 filler statements about the order of the months were constructed by simply reversing the relation between the months in the target statements. All fillers were false. Filler question sets were randomly inserted among the target questions to prevent subjects from deducing the structure of the experiment.

*Procedure:*

Participants were tested individually, in English, in an E-Prime (Psychology Software Tools, Inc., 2002) experiment. Questions were presented on a computer screen one at a time for a maximum of 5 seconds, following Boroditsky (2001). Subjects were instructed to respond T (for true) or F (for false) as quickly as possible by pressing two adjacent keys on a keyboard that had these letters taped onto them. In the practice round, the 5 second response deadline was enforced, and the experimenter encouraged subject to ask any questions they might have. During the experimental round, consultation was prohibited. Response times were measured and recorded by the E-Prime program. Participants received feedback only during the practice round.

*Predictions:*

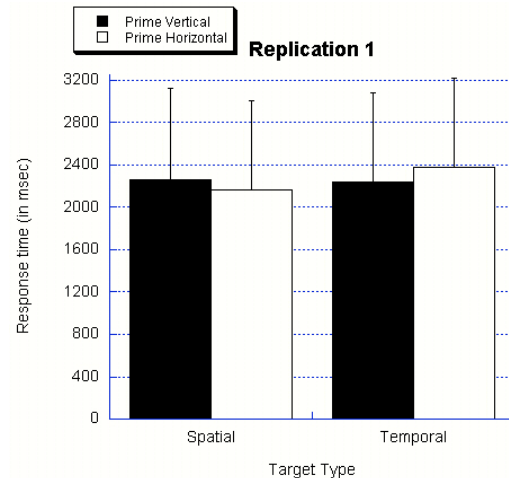
We predict that, as in Boroditsky (2001), English speakers will reliably be able to answer correctly questions about temporal (both the spatial and purely temporal) order faster after horizontal spatial primes than after vertical spatial primes.

**Results:**

A  $2_{\text{prime orientation}} \times 2_{\text{target type}}$  repeated-measures ANOVA was run on the data from the experiment. As the chart below shows, there was almost no difference between

subject response times to targets when the primes were vertically-oriented (2250.2 msec) and when the primes were horizontally-oriented (2271.2 msec),  $F(1, 21) = .48$ ,  $p = .4975$ . Further, there was no effect of target type on target response time (2213.63 msec for spatial targets, 2307.67 msec for temporal targets),  $F(1, 21) = .77$ ,  $p = .3890$  and no target type-prime orientation interaction,  $F(1, 21) = 1.97$ ,  $p = .1753$ . These results contradict the results in Boroditsky (2001), where English speakers answered the targets in 2128 msec after horizontal primes and in 2300 msec after vertical primes.

In both Boroditsky (2001) and the current study, only responses to target temporal questions were analyzed. Further, only those target trials which subjects got entirely correct (i.e., in which they got both primes and the target right) were analyzed. There was a 20.74% error rate in our study, which differs drastically from the error rate in Boroditsky's run (7.1%). Error rates did not differ by target type (21.6% for spatial targets and 19.9% for temporal targets. Error rates did, however, differ by prime orientation (26.1% after horizontal primes and 15.3% after vertical primes),  $\chi^2(1, N = 704) = 12.48$ ,  $p < .0005$ . When error rates were calculated looking only at responses to primes in a trial, they were found to differ significantly by prime orientation (18.47% for horizontal primes and 5.40% for vertical primes),  $\chi^2(1, N = 704) = 28.60$ ,  $p < .0005$ , but not by target type (11.08% preceding spatial targets and 12.78% preceding temporal targets),  $\chi^2(1, N = 704) = .49$ ,  $p > .25$ . When error rates were calculated looking only at responses to targets in a trial, they were found to not differ



significantly by prime orientation (10.51% after horizontal primes and 10.80% after vertical primes),  $\chi^2(1, N = 704) = .01, p > .25$ , or by target type (12.22% for spatial targets and 9.09% for temporal targets),  $\chi^2(1, N = 704) = 1.81, p > .25$ .

**Discussion:**

The results of this experiment do not meet our predictions: subjects did not respond to temporal ordering questions reliably faster after horizontal primes than vertical primes. One potential reason for this fact could be that, as discussed above, “time” is usually discussed as a directional entity. Our prime stimuli, however, made use of static relationships between two objects that are not inherently “predisposed” to movement or direction. Further analysis of Boroditsky’s original stimuli revealed they had this quality. The worms used for the horizontal primes are animals, and thus predisposed to movement, and the “balls” differed crucially from the circles in our experiment in that they were described as “balls” that “float” above or below each other, whereas we called them “circles” and used only the stative-descriptive *be*, not a verb of motion, like *float*.

Subjects made significantly more errors in trials with horizontal stimuli than trials with vertical stimuli, regardless of target type. This is another aspect in which our results differ from Boroditsky’s. That the differences in error rates come solely from the horizontal primes suggests that there is something fundamentally different about our stimuli, compared to Boroditsky’s. Why else would subjects have so much difficulty answering our questions when they did not share this difficulty with Boroditsky’s stimuli?

To see if, indeed, these differences in the disposition to motion were crucial to get the effect of spatial reasoning on temporal reasoning, we reran Experiment 1a with Boroditsky's original stimuli that were more conducive to motion.

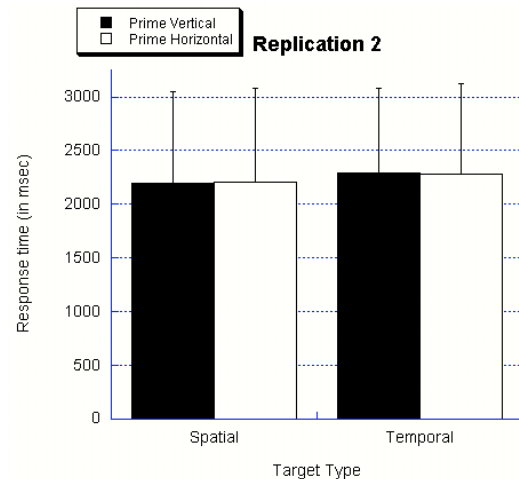
### 2.1b Experiment 1b:

#### Methods:

Methods in Experiment 1b were identical to those in Experiment 1a, except that the spatial primes were exactly the same as those used in Boroditsky (2001) and were much more conducive to motion, as discussed above. The horizontal primes were composed only of screens with different colored worms on them, with "ahead" meaning left and right equally often. Further, half of the vertical primes used the "X floats above Y" construction, the other half used the "X floats below Y" construction (see Figure 3).

#### Results:

As the graph at right shows, subjects still did not answer temporal questions faster after horizontal primes (2243.2 msec) than after vertical primes (2246.3 msec),  $F(1, 19) = .25$ ,  $p = .6222$ . Also, as in the first run of the experiment, there was no effect of target type,  $F(1, 19) = 1.42$ ,  $p = .2480$ , nor a target type-prime orientation interaction,  $F(1, 19) = .04$ ,  $p = .8386$ .



There was a 17.03% error rate for subjects overall, which represents an improvement from the first run. Error rates did not differ by prime orientation (16.56% after horizontal primes and 17.5% after vertical primes),  $\chi^2(1, N = 640) = .10$ ,  $p > .25$ .

However, there was a significant difference in error rates by target type (20.00% for spatial targets and 14.06% for temporal),  $\chi^2(1, N = 640) = 3.99, p < .05$ . When error rates were calculated looking only at responses to targets in a trial, they were found to differ significantly by target type (14.06% for spatial targets and 8.44% for temporal targets),  $\chi^2(1, N = 640) = 5.07, p < .05$  but not by prime orientation (11.25% following horizontal primes and 11.25% following vertical primes),  $\chi^2(1, N = 640) = 0, p > .25$ . When error rates were calculated looking only at responses to primes in a trial, they were found to not differ significantly by prime orientation (6.88% for horizontal primes and 8.44% for vertical primes),  $\chi^2(1, N = 640) = .55, p > .25$ , or by target type (7.50% preceding spatial targets and 7.81% preceding temporal targets),  $\chi^2(1, N = 640) = .02, p > .25$ .

### **Discussion:**

Our results still do not match those of Boroditsky (2001). Since we are using the exact same stimuli, this is a very puzzling result, and the reasons for it are unclear. One point of interest in the results is that whereas Boroditsky's subjects had a 7.1% error rate, our subjects had an error rate of 17.03%. We do not know why our subjects have such a high error rate. We also do not know why there is so much more spread in our response times than in Boroditsky's. It is possible that our subjects are so concerned with getting each question right within the short time limit that, paradoxically, their performance is suffering because their anxiety slows their processing of the temporal and spatial relationships. To see if subject anxiety is responsible for the failure to replicate Boroditsky's results, and to try to elicit the effect one more time, we ran a third replication of Boroditsky's experiment in which we encouraged subjects to be as relaxed as possible during the experiment and not to be as concerned about being accurate.

## 2.1c Experiment 1c:

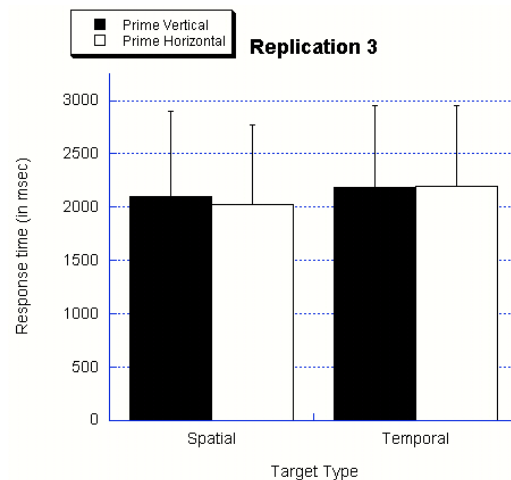
### Methods:

Methods in Experiment 1c were the same as in Experiments 1a and 1b, except subjects completed two practice sessions, the first composed of two trials without a time limit, the second composed of 6 trials. The second, untimed practice round was added before the original practice round to allow subjects more time to get familiar with the set up of the experiment and the types of questions that would be asked of them. Also, the instructions to the experiment were changed to specifically encourage subjects to make speed a priority and not to worry too much about getting each question right. The experimenter also verbally encouraged them to be quick and not to be over-concerned with accuracy.

### Results:

Again, subjects did not respond faster after horizontal primes (2114.4 msec) than after vertical primes (21265.5 msec),  $F(1, 23) = .55, p = .4663$ . There was a nearly significant effect of target type on response time,  $F(1, 23) = 4.00, p = .0574$ . There was no target type-prime orientation interaction,  $F(1, 23) = 3.25, p = .0844$ .

The error rate for this run was 17.32%. Error rates did not differ by prime orientation (15.36% after vertical primes and 19.27% after horizontal primes),  $\chi^2(1, N = 384) = 1.10, p > .25$ , or target type (18.75% for spatial targets and 17.71% for temporal





targets),  $\eta^2(1, N = 384) = 2.04, p > .15$ . There was, however, a significant effect of prime orientation on error rate when calculating the error rate based solely on responses to targets (12.76% after horizontal primes and 7.81% after vertical primes),  $\eta^2(1, N = 384) = 5.09, p < .05$ .

### **Discussion:**

The results of our third replication of Boroditsky's experiment still do not match hers. Further, the only significant effect on error rate is in the opposite direction of what would be expected if indeed there were a relativistic priming effect of spatial computation on temporal computational. If English speakers' representation of time is horizontal (because of the linguistic treatment of time in English), then one would expect that reasoning about horizontal spatial relationships would facilitate reasoning about temporal relationships. This facilitation would be manifested in both the response time *and* the accuracy to the temporal questions, response times getting faster and accuracy improving. Our results show no effect on response time and an apparent detrimental effect of horizontal spatial reasoning on temporal reasoning. There was a nearly significant effect of target type on response time, but since the targets were the same in this experiment as the last experiment there is no reason to believe that this is a result of anything but chance variation, as opposed to an assumption that if we could relax subjects more we would get a significant effect. Even if we do take this trend toward significance as such evidence, the effect reported in Boroditsky (2001) must be fragile in deed if it requires an extreme low stress level in subjects. In light of these facts, it seems natural to conclude that there is no relativistic effect on temporal computation and that

the results in Boroditsky (2001) are a fluke. Other experimenters, however, have been able to replicate her results (Boroditsky, pers. comm.).

Recent communications with Boroditsky have revealed some potentially important misinterpretations of the design of the original experiment. For example, filler trials in Boroditsky (2001) were composed only of 2 screens, and the type (spatial or temporal) and truth value of each screen was randomly determined, potentially resulting in some filler trials that were composed of two false temporal screens. This arrangement is impossible in the current study: all filler trials were composed of 3 screens, the first two of which were always spatial, the third always temporal. Further, the 5 second response time was not strictly enforced during the experiment in Boroditsky (2001); each screen required a response to advance, and all responses that took longer than 5 seconds were later filtered out. In the current study, if subjects did not respond within 5 seconds of a screen's appearance, the screen automatically moved on. Perhaps such a strict enforcement of the time limit might drastically affect subjects' performance in the experiment: they might either wait for the time limit to lapse on screens they have some initial difficulty in answering, or the constant reminder of time pressure might cause subjects excessive stress (Boroditsky (pers. comm.)). That the  $F$  values changed so much from the second run of this experiment ( $F(1, 19) = .04, p = .8386$ ) to third ( $F(1, 23) = 3.25, p = .0844$ ) lends some support to this theory: when subjects are encouraged to be relaxed and given an untimed introduction to the experiment, the interaction between target type and prime orientation becomes nearly significant, when it was decidedly not significant without this "relaxed" introduction. We do not know at this time exactly how these differences could have affected subjects, but we do recognize them as capable of

doing so. Further, it is also unclear what it means for the effect to be so fragile that small changes in the experiment or the environment can destroy it.

## **2.2 Experiment 2:**

Boroditsky (2001) found a priming effect of spatial language on temporal reasoning. However, we do not know how much influence on temporal reasoning the spatial language has. Boroditsky (2000) reports work that addresses this question. She noted that there are two types of metaphors people use to talk about time: people talk of time either as if it is a substance they move through (“*We’re coming up* on New Year’s Eve.”), called the ego-moving perspective, or as if it moves past them (“New Year’s is almost *upon us*.”), called the time-moving perspective. In her study, Boroditsky tested to see if thinking about spatial relationships from an ego-moving perspective led people to think of temporal relationships from an ego moving perspective and if thinking spatially from an object-moving perspective led people to think about time from the time-moving perspective; that is, she tested for cross-domain consistency of perspectives. In her study, she found cross-domain consistency of perspectives from the spatial domain to the temporal domain, but not from the temporal domain to the spatial domain, both in off-line and online tasks (Boroditsky, 2000).

The next experiment in the current study attempts to discover more about the strength of this effect of spatial language on temporal cognition by asking if, because of the spatial metaphors used to talk about time, space and time have similar representational structures. If so, then temporal computation should interfere with spatial memory in the same way that spatial computation does. It is also likely that spatial computation will interfere with temporal memory in the same way that temporal

computation does if the two domains share representations. Though Boroditsky (2000) found that there was only a one-way priming effect from thinking about spatial relationships to thinking about temporal relationships, she was testing for inter-domain consistency in metaphor use (ego-moving versus time-moving), not for interference based upon shared representations. If temporal and spatial cognition interfere with each other in ways that the mathematical or object recognition domains do not, then the concepts of time and space make use of shared representations in the mind. The only reason for these shared representations would be the linguistic treatment of time. Experiment 2 was designed to test for these interference effects.

### **Methods:**

#### *Participants:*

Participants were 48 students at Swarthmore College, who had learned English before age 5. Compensation for participation in the experiment was \$5.

#### *Procedure and Design:*

Subjects were shown a series of images and sentences on a computer screen. Each subject ran 16 trials, and each trial was composed of a target screen, 3 distracter screens, and a question screen.

Table 3: Design of Experiment 2

<b>Target:</b> Spatial	<b>Target:</b> Spatial	<b>Target:</b> Spatial
<b>Distracters:</b> Spatial	<b>Distracters:</b> Temporal	<b>Distracters:</b> Other
<b>Target:</b> Temporal	<b>Target:</b> Temporal	<b>Target:</b> Temporal
<b>Distracters:</b> Spatial	<b>Distracters:</b> Temporal	<b>Distracters:</b> Other

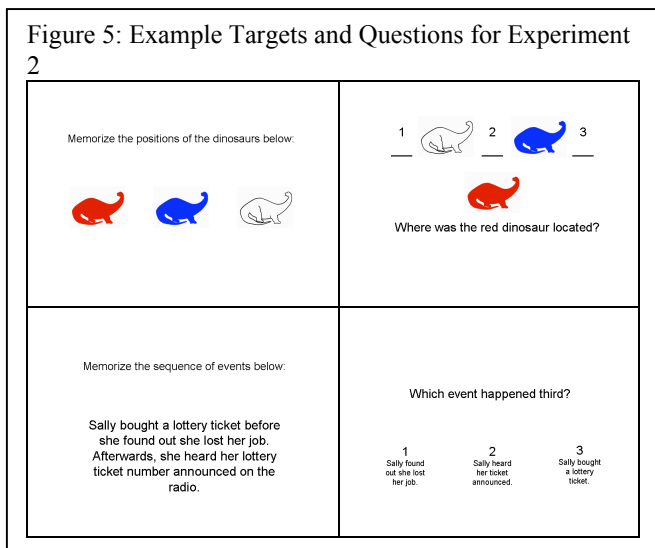
Subjects were instructed to memorize the information on the target screen, shown for 15 seconds, and then to answer the True/False questions on the distracter screens within a 5 second deadline before answering the question screen within a 10 second deadline.

Targets and distracters were combined to create 6 experimental conditions (see Table 3),

which can be grouped into two halves: the spatial-first half, in which subjects had to memorize the position of animals and answer one of 3 types of distracter questions, and the temporal-first half, in which subjects had to memorize a sequence of events before the distracters. Each participant was run in only one condition to avoid fatigue in answering 96 trials, resulting in a fully between-subjects design. Subjects were all run individually, and each was given a practice round using letters and their spatial relationships to familiarize themselves with the structure of the experiment.

*Materials:*

A set of 16 target and question screen pairs and 48 distracter screens was constructed.

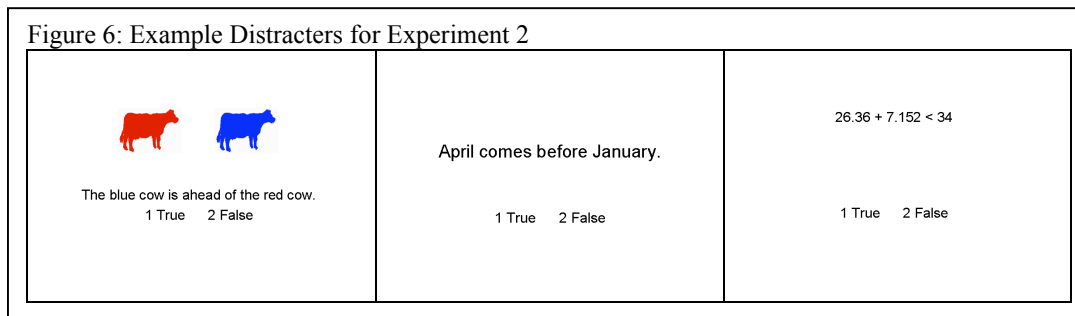


*Targets and Questions.*

Targets were of 2 types, either an array of three animals of different colors or a short paragraph describing a sequence of events. Questions were also of two types, to avoid the development of response strategies by subjects. In the spatial-

first group, the questions asked the subject to locate which animal was in a given spot in the line or to indicate which spot in the line a given animal filled. In the temporal-first group, the questions asked subjects to locate when in the sequence a given event occurred or to indicate which event occurred in a given point in the sequence. (See Figure 5).

*Distracters.* Distracters were of 3 types: spatial, temporal, or math. Spatial distracters were pictures of 2 animals with a statement underneath describing their relative positions. Temporal distracters were sentences describing the temporal order of 2 months in a year. Math distracters were arithmetic inequalities. All distracters were designed to elicit a True or False response from the subject. Spatial distracters equally often used *ahead* and *behind*, temporal distracters equally often used *before* and *after*, and math distracters equally often used *greater than* and *less than*, and drew equally from addition, subtraction, multiplication, and division problems. These problems were then randomly distributed among the target-question pairs (see Figure 6 for examples).



*Predictions:*

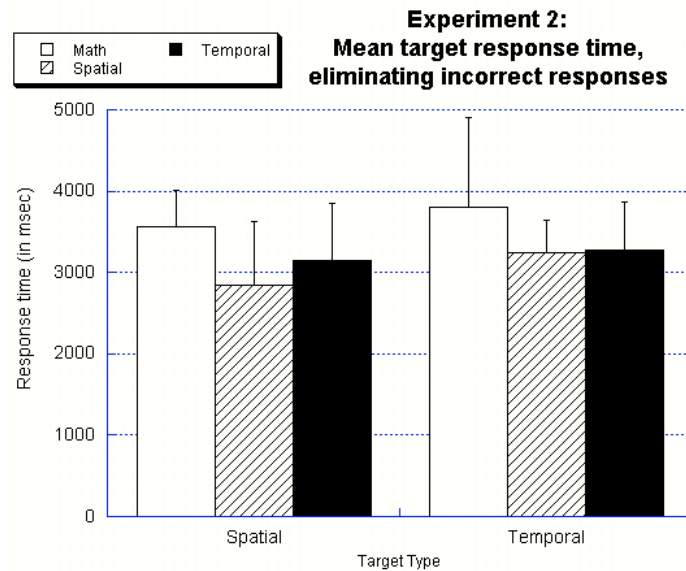
If the Linguistic Relativity Hypothesis is correct and time has a representational structure similar to space because of the way people talk about it, then subjects' accuracy and response times should pattern similarly in the spatial-spatial and spatial-temporal conditions, as opposed to the spatial-math condition. There should also be a similar interference pattern in the temporal-spatial and temporal-temporal conditions, as opposed to the temporal-math condition.

**Results:**

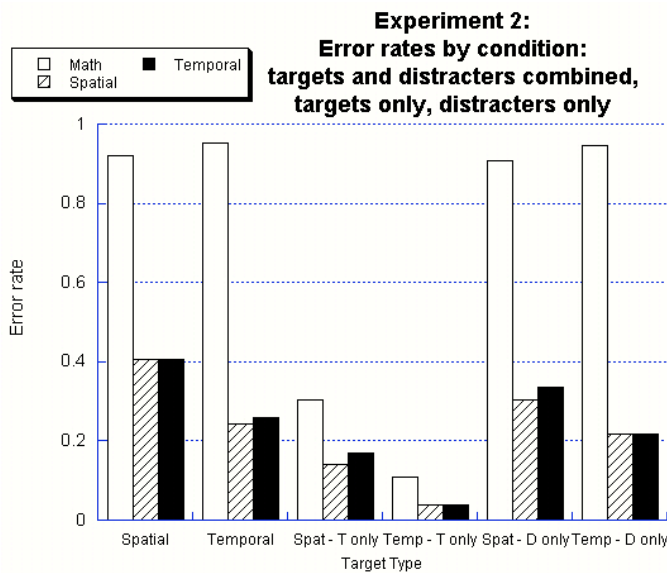
Only response times to target questions were analyzed. We did not follow the Boroditsky's practice of dropping trials in which subjects answered any of the distracters or the target incorrectly for the response time analysis because that would leave 10 or fewer observations before collapsing over items in some cells (10 in the spatial-math condition and 6 in the temporal-math condition). However, we did drop observations in which subjects did not answer the target question correctly.

In both the spatial and temporal superconditions, reaction times were longer after math distracters (spatial: 3995.6 msec; temporal: 3891.1 msec) than after either the spatial

distracters (spatial: 2931.6 msec; temporal: 3255.9 msec) or temporal distracters (spatial: 3292.3 msec; temporal: 3295.4 msec). A  $2_{\text{target type}} \times 3_{\text{distracter type}}$  ANOVA was run on the mean response time per condition. A significant main effect of distracter type was found ( $F(2, 1) = 7.33, p = .0019$ ), but there was no effect of target type ( $F(1, 2) = .027, p = .6066$ ) and no target-distracter interaction ( $F(2, 42) = .48, p = .6231$ ). These results did not change when trials with incorrect responses to the targets were eliminated: for distracter type  $F(2, 1) = 3.58, p = .0366$ ; for target type  $F(1, 2) = 2.54, p = .2220$ ; for the interaction  $F(2, 42) = .13, p = .8777$ .



Error rates differed dramatically by distracter type (math: 93.75%; spatial: 32.42%; temporal: 33.20%),  $\chi^2(2, N = 768) = 254.53, p < .0005$ , and by target type (spatial: 86.72%; temporal: 72.66%),  $\chi^2(1, N = 768) = 6.78, p < .02$ . When error rates were calculated using only responses to distracter questions, they were found to differ by



$p < .0005$ , but not by target type

(spatial: 77.34%; temporal:

69.14%),  $\chi^2(1, N = 768) = 2.30$ ,

$p > .10$ . When error rates were

calculated using only responses to

target questions, they were found to

differ by both distracter type (math:

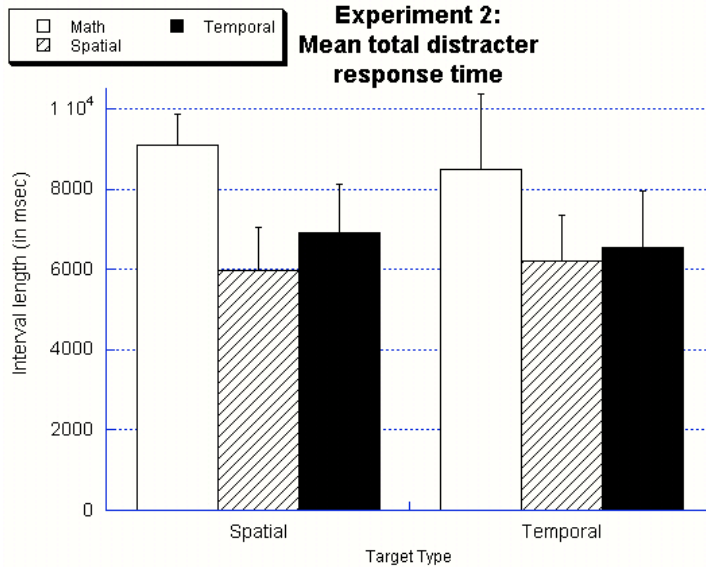
20.70%; spatial: 8.98%; temporal:

9.77%),  $\chi^2(2, N = 768) = 19.24, p < .0005$ , and target type (spatial: 30.08%; temporal:

9.38%),  $\chi^2(1, N = 768) = 32.02, p < .0005$ .

Though there was an enforced time limit in this experiment, it is possible that subjects were reliably slower to respond to certain types of distracters than others, which would make total distracter response time a confounding variable in this experiment. There were no total distracter response times that are greater than 3 standard deviations above the mean in Experiment 2, so there were no observations eliminated by that criterion. Further, eliminating observations in which the subject had answered at least 1 distracter incorrectly would have left only 12 observations in the spatial-math condition and 7 in the temporal math condition before collapsing over items, so we did not





eliminate observations based on that criterion either. A  $2_{\text{target type}} \times 3_{\text{distracter type}}$  ANOVA revealed a significant effect of distracter type on total distracter response time,  $F(2, 1) = 54.07, p = .0000$ , with no other significant effects. Thus, it appears that total

distracter response time is a confounding variable for target accuracy in Experiment 2.

### Discussion:

In both halves of the experiment, response times were unexpectedly high and accuracy unexpectedly low in the math condition. Response times were longer and accuracy scores lower in the math condition than in either the spatial-spatial condition or the temporal-temporal condition, though it seems reasonable to expect that such same-domain reasoning should result in the strongest interference effect.

There are two possible ways to explain these unexpected results. The first is to interpret them as evidence for a facilitation effect of same-domain reasoning on memory—that is, to claim that reasoning about the same type of thing as that which one is trying to remember aids memory in that domain. This is counterintuitive because one expects that reuse of the same representations should result in an “overwriting” of the information already in play in those representations.

The counterintuitive nature of such a facilitation effect notwithstanding, the experiment done by Boroditsky (2000) mentioned in the introduction to the second experiment supports the idea that one exists. As mentioned previously, there are two perspectives used to discuss objects moving in space with respect to the speaker: the ego-moving perspective, in which the speaker is seen as moving and the objects as stationary, and the object-moving perspective, in which this is reversed. An example of the ego-moving perspective in language is “I’m coming to the door”; an example of the object-moving perspective is “The car is getting closer to me”. Similarly, there are two perspectives used to talk about temporal relationships and progress: the ego-moving perspective, in which the speaker is seen as moving through time, and the time-moving perspective, in which events are seen as moving toward (or away from) the speaker. An example of the ego-moving perspective is “We’re coming up on New Year’s Eve”; an example of the time-moving is “New Year’s Eve is almost here”.

In her experiment, Boroditsky (2000) showed subjects 2 spatial primes that picked out either the ego-moving perspective or the

Table 4: Design of the experiment in Boroditsky (2000)

<b>Prime:</b> Spatial; Ego	<b>Prime:</b> Spatial; Object	<b>Prime:</b> Temporal; Ego	<b>Prime:</b> Temporal; Time
<b>Target:</b> Spatial; Ego	<b>Target:</b> Spatial; Ego	<b>Target:</b> Spatial; Ego	<b>Target:</b> Spatial; Ego
<b>Prime:</b> Spatial; Ego	<b>Prime:</b> Spatial; Object	<b>Prime:</b> Temporal; Ego	<b>Prime:</b> Temporal; Time
<b>Target:</b> Spatial; Object	<b>Target:</b> Spatial; Object	<b>Target:</b> Spatial; Object	<b>Target:</b> Spatial; Object
<b>Prime:</b> Spatial; Ego	<b>Prime:</b> Spatial; Object	<b>Prime:</b> Temporal; Ego	<b>Prime:</b> Temporal; Time
<b>Target:</b> Temporal; Ego	<b>Target:</b> Temporal; Ego	<b>Target:</b> Temporal; Ego	<b>Target:</b> Temporal; Ego
<b>Prime:</b> Spatial; Ego	<b>Prime:</b> Spatial; Object	<b>Prime:</b> Temporal; Ego	<b>Prime:</b> Temporal; Time
<b>Target:</b> Temporal; Time	<b>Target:</b> Temporal; Time	<b>Target:</b> Temporal; Time	<b>Target:</b> Temporal; Time

There were sixteen conditions, crossing prime type, target type, and schema type. ‘Ego’ indicates those scenarios that used the ego-moving schema; ‘Object’, the object-moving schema; ‘Time’, the time-moving schema.

object-moving perspective or 2 temporal primes that picked out either the ego-moving

perspective or the time-moving perspective and asked the subjects to indicate whether the sentences on them were true or false. Next, she showed subjects target questions that were either spatial or temporal, and equally often used the ego-moving and object-moving perspective (for spatial targets) and the ego-moving and time-moving perspective (for temporal targets). Each target appeared twice in the experiment, once when the schemas of the primes (ego-moving or object/time-moving) were consistent with the schema used in the target, and once when they were inconsistent. Table 4 lays out all 16 conditions in the experiment.

The critical measure in the experiment was the effect of schema consistency on response time to the target. Boroditsky calculated the “consistency bias” of subjects by subtracting the response time for a given target in the consistently primed condition from the response time in the inconsistently primed condition. She found a reliable effect of schema consistency (that is, a reliable difference between response time for the inconsistently primed targets and the consistently primed targets) when primes and targets were of the same domain type (that is, when primes and targets were both spatial or both temporal), and also when the primes were spatial and the targets temporal, but there was no reliable effect when the primes were temporal and the targets spatial. To put it another way, spatial computation primes temporal computation, but temporal computation does not prime spatial computation, and both domains prime same-domain computation.

This conclusion is relevant to the current study in that it seems highly plausible that same-domain computation could, indeed, facilitate same-domain memory. Assuming that this is true, the results of Experiment 2 could be viewed as confirmation of

the Linguistic Relativity Hypothesis: the response times and error rates in the temporal and spatial conditions pattern together because the temporal and spatial domains make use of shared representations. Subjects do comparatively worse in the math condition because switching between domains, from the spatial/temporal domain to the mathematical domain, and then back, is a highly-difficult task that overburdens the mind.

The limitation of this explanation comes from the fact that in Boroditsky (2000) there was an asymmetrical priming relationship: space primed time, but time did not prime space. This asymmetry should manifest itself in the current study through higher error rates on the distracters and longer intervals between the memorization screen and target recall question in the temporal-spatial condition than in the other conditions (excluding the math distracter conditions because they are not accounted for in Boroditsky (2000)). The same-domain priming effect reported in Boroditsky (2000) that should make the error rates smaller and interval length shorter in the spatial-spatial and temporal-temporal conditions. Error rates for the distracters in the temporal-spatial condition are equal to those in the temporal-temporal condition, contrary to this prediction. Further, the asymmetry in priming effects should cause larger error rates and longer response times on targets in the spatial-temporal condition, since this is another instance in which spatial questions follow temporal questions. However, there are no reliable differences between the targets in accuracy or response time between the spatial-spatial and spatial-temporal conditions. So it seems that the asymmetrical priming effect reported in Boroditsky (2000) might not be able to account for the data in the current study.

Another possible explanation for the unexpected results in Experiment 2 holds that the arithmetic problems in the math condition may have been too difficult to answer in such a short time window (5 seconds) and that their difficulty was very distressing to subjects, causing them to perform badly. The marked differences in error rate by distracter type support this view. For subjects to perform so badly (i.e., to get 90% of the questions wrong) the distracter questions must have been vastly more difficult in the math condition than in either the spatial or temporal condition. This tremendous difficulty could have overwhelmed subjects and caused a general computational failure, which would account for the longer response times and lower accuracy in the math condition.

To decide between these possible interpretations, another experiment in which we adjust the level of difficulty of the arithmetic problems to match the difficulty of the other distracters is necessary. Further, comparison of all of these load conditions with a low load condition in which subjects have to answer very simple questions would also be very informative. Even if subjects response times to the arithmetic questions are slower than in the spatial or temporal distracter conditions when the difficulty of each condition is more evenly matched, performance during the low load condition provides another basis for comparison, allowing us to conclude that mathematical reasoning is just inherently more difficult and requires more mental resources than either spatial or temporal reasoning. Experiment 3 was created to help disambiguate the results of Experiment 2 in this way.

### **2.3 Experiment 3:**

#### **Methods:**

##### *Participants:*

Participants were 16 students at Swarthmore College recruited by placing signs around the campus. Compensation for participation in the study was \$5.

*Procedure and Design:*

Participants were tested individually in an E-Prime experiment. Subjects completed 40 trials in an interference/recall experiment, comprising 4 conditions. Each trial was composed of 5 screens. The first screen in each trial instructed subjects to memorize the order of the animals or sequence of events on it. The next three screens each had one of each of the following types of distracters: spatial questions, temporal questions, mathematical questions, and shape-matching questions. All distracter questions required either a true/false or yes/no responses. The final screen in the set of 5 asked where an animal had been on the initial screen or when an event occurred in the initial presentation and gave 3 options. All screens stayed up until the subject indicated he or she was ready to proceed, either by pressing the spacebar (for the memorization screens) or by responding to questions (for the distracter and target question screens). Additionally, subjects were given an opportunity to rest as long as they liked between trials. The order of trials was newly randomized for each subject. The practice round consisted of 4 trials, each an exemplar of the four conditions in the experiment. Subjects received reaction time and accuracy feedback only during the practice rounds. The experiment had a partially within-subjects design, each subject completing either all 4 spatial memory conditions (spatial memorization and each of spatial, temporal, math, and low load distracters) or all 4 temporal memory conditions (temporal memorization and each of spatial, temporal, math, and low load distracters).

*Materials:*

Materials were very similar to those used in Experiment 2. The chief difference was the inclusion of the low load (shape-matching) distracters and the simplification of the math problems.

*Targets and Questions.* Additional targets and questions were created to make a total of 40 pairs of spatial targets and questions and 40 pairs of temporal targets and questions (plus four pairs of each type for the practice rounds), each appearing once during the experiment.

*Distracters.* Distracters in both the temporal and spatial conditions were a randomly selected set of 30 from the 48 used in Experiment 3. In the mathematical condition, a new set of 30 equations using only addition and subtraction was constructed, of which 15 were true and fifteen false. These problems were then randomly distributed among the ten trials that compose the mathematical condition. For the low load condition, a set of 30 simple line drawings asking if the shapes on them matched was constructed using autosshapes from Microsoft PowerPoint. These drawings were then distributed over the 10 trials of the low load condition.

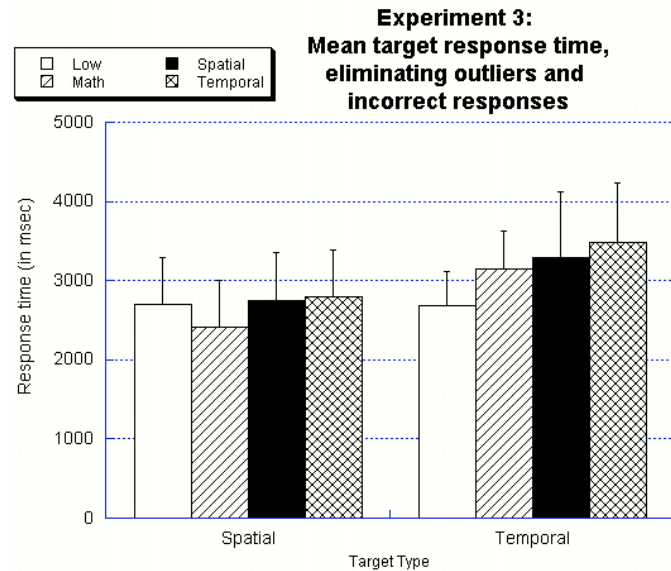
*Predictions:*

We predict that, because of the lesser degree of difficulty in the mathematical condition, subjects' response times to question screens will be faster after mathematical distracters than after either spatial distracters, when the initial array is spatial, or temporal distracters, when the initial display is temporal. We further predict that they will be fastest of all after low load distracters. If the Linguistic Relativity Hypothesis is correct, then we predict that response times in the spatial and temporal conditions, regardless of

the initial display, will pattern together when compared to the mathematical or low load conditions.

### Results:

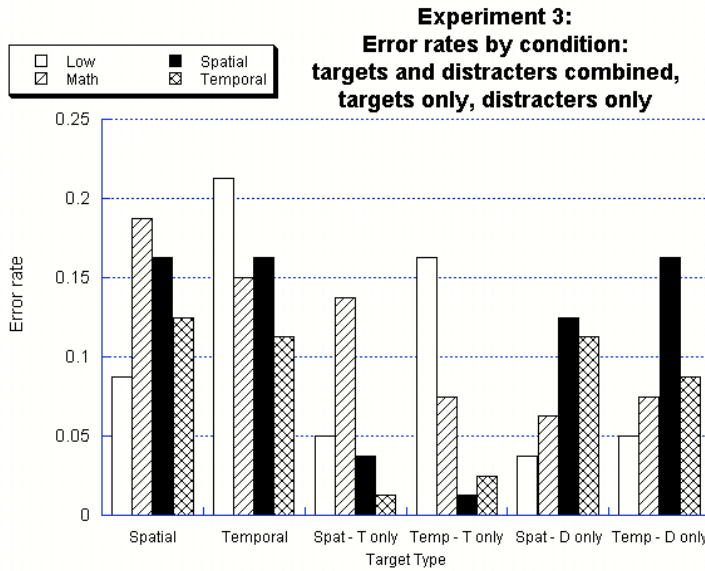
Unlike in the other experiments in the study, there was no enforced time limit for responses to either targets or distracters. As a result, all observations in which response times to the trials were more than 3 standard deviations above the overall mean were omitted from the



response time analysis. A  $2_{\text{target type}} \times 4_{\text{distracter type}}$  ANOVA was run on the data collected in the above experiment. This time, a significant main effect of target type on response time was found ( $F(1, 3) = 10.09, p = .0024$ ), but there was no effect of distracter type ( $F(3, 1) = 1.80, p = .1579$ ) and, again, no target-distracter interaction ( $F(3, 56) = 1.23, p = .3061$ ). These results didn't change even when all observations were included: there was a significant main effect of target type on response time ( $F(1, 3) = 5.69, p = .0204$ ), but there was no effect of distracter type ( $F(3, 1) = .89, p = .4522$ ) and, again, no target-distracter interaction ( $F(3, 56) = .71, p = .5479$ ).

Error rates did not differ by distracter type (low: 8.75%; math: 16.88%; spatial: 16.25%; temporal: 11.88%),  $\chi^2(3, N = 320) = 1.875, p > .25$ , or target type (spatial:





14.06%; temporal: 15.94%),  
 $\chi^2(1, N = 320) = .44, p > .25^9$ .

When error rates were calculated based solely on responses to the distracters, they were found to significantly differ by distracter type (low: 4.38%; math: 6.88%; spatial:

14.38%; temporal: 10.00%),  $\chi^2(3, N = 320) = 11.00, p < .02$ , but not by target type (spatial: 8.44%; temporal: 9.38%),  $\chi^2(1, N = 320) = .17, p > .25$ . When error rates were calculated based solely on responses to the targets, they were also found to significantly differ by distracter type (low: 10.63%; math: 10.63%; spatial: 2.50%; temporal: 1.88%),  $\chi^2(3, N = 320) = 19.05, p < .0005$ , but not by target type (spatial: 5.94%; temporal: 6.88%),  $\chi^2(1, N = 320) = .23, p > .25$ .

Since there were no time limits for responses in this experiment, there is a chance for much greater variability in the total distracter response time. If there are differences among the various conditions on the total distracter response time, the length of the interval could be a confounding variable in the analysis. To address this concern a  $2_{\text{target type}} \times 4_{\text{distracter type}}$  ANOVA was run on mean total distracter response time<sup>10</sup>, eliminating total distracter response times greater than 3 standard deviations above the mean for all distracter types. A very significant main effect of distracter type on total distracter

<sup>9</sup> The Experiment 2 and Experiment 3 graphs for error rates are on different scales because they differ so drastically. If the Experiment 3 error rate graph were on the same scale as the Experiment 2 graph, some of the differences between conditions would be extremely difficult to see.

response time was found,  $F(3, 1) = 31.04, p = .0000$ , with no other significant effects.

These results did not change when all total distracter response times were included: for

distracter type,  $F(3, 1) =$

22.14,  $p = .0000$ . Thus it

appears that total distracter

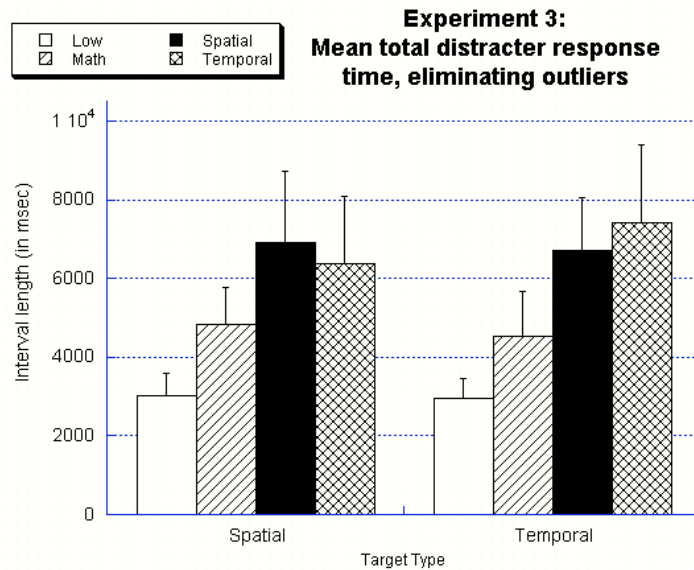
response time is a

confounding variable for

target accuracy in

Experiment 3, as it was in

Experiment 2.



The differences in the design of the experiments in Experiment 2 and Experiment 3 allow us to test how an enforced time limit might affect subjects in a response time experiment, suggested as a potential reason for our current inability to reproduce Boroditsky's (2001) results in Experiment 1. Though we cannot directly compare response times to the distracter questions in the two experiments because they were a between-subjects variable in Experiment 2 and a within-subjects variable in Experiment 3, we can compare response times to the targets, which were a between-subjects variable in both studies. A  $2_{\text{target type}} \times 2_{\text{time constraint}}$  ANOVA run on the mean correct response times to target questions in the spatial and temporal distracter conditions, eliminating outliers in Experiment 3, indicates an effect of target type on response time,  $F(1, 1) = 5.36, p = .0253$ , but no effect of time constraint on response time, and no target type-time constraint interaction (both  $ps > .30$ ). However, there is a significant effect of time

<sup>10</sup> To keep the results of this ANOVA more comparable to the one conducted on the intervals in Experiment 2, we did not eliminate trials in which subjects answered one or more distracters incorrectly.

constraint on accuracy,  $F(1, 1) = 20.03, p = .0001$ . Also, there was a significant effect of target type on accuracy,  $F(1, 1) = 10.65, p = .0021$ , and a target type-time constraint interaction,  $F(1, 1) = 10.65, p = .0021$ .

### **Discussion:**

The results do not indicate that temporal and spatial computation share the same mental representations; subjects did not take longer to answer target recall questions based on the type of distracter computation they had performed between the presentation of the target and its recall question.

A point of interest in the data is that the pattern of errors changes depending whether we look at error rates for targets alone or error rates for distracters alone. That is, subjects make more errors on targets regardless of target type in the low and math distracter conditions (low: 10.63%; math: 10.63%; spatial: 2.50%; temporal: 1.88%),  $\chi^2(3, N = 320) = 19.05, p < .0005$ , and more errors on distracters regardless of target type in the spatial and temporal distracter conditions (low: 4.38%; math: 6.88%; spatial: 14.38%; temporal: 10.00%),  $\chi^2(3, N = 320) = 11.00, p < .02$ . One potential reason for this reversal in patterning is that, though the low and math distracters are easier by themselves, they interfere more with temporal and spatial memory because of the domain shift. This explanation is in line with the potential facilitation effect of same-domain processing that could be the explanation for the patterning of the error rates in Experiment 2.

The results of Boroditsky (2000) might lend some credence to explanation, and allow us to interpret the results as in support of the Linguistic Relativity Hypothesis because space and time use shared mental representations. However, this seems less

plausible as an explanation, again, because of the asymmetrical nature of the priming relationship. As in Experiment 2, we expect higher error rates on distracters and longer intervals in the temporal-spatial condition and higher error rates and longer response times to targets in the spatial-temporal condition. Though the error rate on distracters in the temporal-spatial condition is much higher than in the temporal-temporal condition, the interval length is shorter in the temporal-spatial condition. Further, the response time to targets in the spatial-temporal condition is not significantly different from the spatial-spatial condition, and the error rate on targets is less in the spatial-temporal condition than in the spatial-spatial condition.

Another potential reason for this patterning of error rates is related to the differences in interval length for the various distracter types. Since subjects are answering questions faster in the low and math conditions, they might be more prone to indicate a response to a target by mistake, to indicate position 1 when they mean position 2. This explanation is more appealing, again, because of the counterintuitive nature of a facilitation effect that is its alternative. Under this explanation, answering the low and math distracters do not cause subjects greater processing difficulty than the spatial or temporal distracters, but rather their relative ease leads to errors in reporting the intended response.

No matter which interpretation of the error rates results we adopt for Experiments 2 and 3, we still cannot decisively conclude that the results do or do not support the Linguistic Relativity Hypothesis because there is the confounding variable of the length of the interval between the end of the memorization screen and the recall question. In both Experiment 2 and Experiment 3, subjects took either significantly longer or shorter

to answer the set of distracter questions in the math distracter condition than in either of the spatial or temporal distracter conditions. It seems reasonable to conclude that the amount of time needed to answer a question is related (positively) to its difficulty. It seems, then, that we have failed to find math distracter questions that are matched for difficulty with the spatial and temporal distracter questions, which is a definite confound in the analysis. Further, upon reflection it seems that it would be best to ensure that the low load distracter interval is as long on average as the spatial and temporal distracter intervals. Since we cannot do this by matching the various distracter types for difficulty, the low load distracters intentionally being easier than the others, perhaps having more low load distracters in a distracter interval would be the best way to accomplish this. That way, the interval length variable would be more nearly controlled and we could truly compare the effects of the various computation types on memory.

### **3. GENERAL DISCUSSION AND CONCLUSION**

The results of 3 experiments investigating the Linguistic Relativity Hypothesis have been inconclusive. The first experiment failed to replicate the results of Boroditsky (2001), which found that English speakers could verify purely temporal target questions faster after answering questions about horizontal spatial relationships than after vertical spatial relationships, while this pattern was reversed for native Mandarin speakers tested in English. This asymmetry in the results of Boroditsky (2001) is taken as evidence for a weaker version of the Linguistic Relativity Hypothesis because the only plausible explanation for the differences in response pattern is the different linguistic treatment of time in English (horizontal) and Mandarin (vertical).

The reasons for the inability to replicate these results for English speakers in the current study are unclear, though there were some differences in structure between the experiment in Boroditsky (2001) and the current study. Most conspicuous of these is that the 5 second response time limit was strictly enforced in the current study during the experiment (the screen with the question on it disappeared after 5 seconds), whereas in Boroditsky (2001) subjects could take as much time as they wanted to answer the questions, response longer than 5 seconds were filtered out after the experiment was completed.

One potential reason for the inability to replicate Boroditsky's results put forward is that the strict enforcement of the time limit increased the subjects' stress level during the experiment and caused their performance to degrade as a result. To test the plausibility of this explanation, the response times to target questions in Experiment 2, which were timed, and Experiment 3, which were untimed were compared. A reliable effect of question type on reaction time was found, but no effect of whether the question had a strictly enforced time limit. Additionally, reliable effects of question type, time limit, and a question type-time limit interaction on accuracy were found.

The question type effects are less relevant to the Boroditsky experiment because the questions are very much different from the Boroditsky experiment and either of Experiment 2 or Experiment 3. However, the reliable effect of time limit on accuracy, though not immediately related to reaction time, might explain the vast discrepancy in error rates from Boroditsky (2001) and the current study, though this effect of time limit on accuracy was not tested for the experiment in Boroditsky (2001). Further, since only

response times to target questions answered correctly were analyzed, accuracy could potentially impact the mean response time.

Because of these effects on accuracy, and because other experimenters have reportedly been able to replicate Boroditsky's (2001) results, we do not interpret our results as a refutation of the effect of language on thought found in Boroditsky (2001). However, we also cannot conclude that our results support the Linguistic Relativity Hypothesis because they clearly do not. What we can conclude is that the effect of language on thought reported in Boroditsky (2001) is very fragile. If the orientation of the spatial metaphor used to discuss time does affect the mental representation of time, it must do so in a very superficial way; clearly, the linguistic treatment of time does not fully determine the representation of it if the effect of the linguistic treatment can vanish under moderate stress.

The results of the second 2 experiments in the current study do not give us much more information about the correctness of the Linguistic Relativity Hypothesis. The confounding variable of total distracter response time in both experiments prevents any interpretation of the differences in response time and accuracy in terms of the types of computation between the memorization screen and the target recall question. However, the results do suggest that there is no asymmetrical priming effect from spatial computation to temporal computation as reported in Boroditsky (2000) occurring. Indeed, our results exhibit evidence against this priming effect.

In light of these facts, we cannot conclude anything about the Linguistic Relativity Hypothesis based on the experiments in this paper. However, we have demonstrated the fragility of the effect in Boroditsky (2001) and Boroditsky (2000).

Further, we have developed a tool used in Experiments 2 and 3 to test for shared representations for the domains of space and time that, when instantiated in a way that eliminates interval length as a confounding variable, will be able to tell us if the Linguistic Relativity Hypothesis is indeed correct.



**References**

- Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Los Angeles: University of California Press.
- Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75, 1-28.
- Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43, 1-22.
- Brown, R. (1976). Reference: In memorial tribute to Eric Lenneberg. *Cognition*, 4, 125-153.
- Brown, R. & Lenneberg, E. (1954). A study in language and cognition. *Journal of Abnormal and Social Psychology*, 49, 454-462.
- Heider, E. & Olivier, D. (1972). The structure of the color space in naming and memory for two languages. *Cognitive Psychology*, 3, 337-354.
- Jeanne, L. (1978). Aspects of Hopi Grammar. In K. Hale & L. Jeanne (Eds.), *Issues of Hopi Linguistics in Theoretical and Typological Perspectives*. New Mexico: University of New Mexico Linguistics Summer Program.
- Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago: The University of Chicago Press.
- Levinson, S.C. (1996). Language and Space. *Annual Review of Anthropology*, 25, 353-382.
- Levinson, S.C. (2001). Covariation between spatial language and cognition, and its implications for language learning. In M. Bowerman & S. C. Levinson (Eds.), *Language Acquisition and Conceptual Development* (pp. 566-588). New York: Cambridge University Press.
- Levinson *et al.* (2002). Returning the tables: Language affect spatial reasoning. *Cognition*, 84, 155-188.
- Li, P. & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition*, 83, 265-294.
- Psychology Software Tools, Inc. (2002). E-Prime (Version 1 and Beta 1) [Computer software]. Pittsburgh, PA: Psychology Software Tools, Inc.

- Whorf (1939). The relation of habitual thought and behavior to language. In J. B. Carroll (Ed.), *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (pp. 134-159). Cambridge, MA: The MIT Press, 1956.
- Whorf (1941). Language, mind, and reality. In J. B. Carroll (Ed.), *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (pp. 134-159). Cambridge, MA: The MIT Press, 1956.
- Whorf (1956a). A linguistic consideration of thinking in primitive communities. In J. B. Carroll (Ed.), *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (pp. 134-159). Cambridge, MA: The MIT Press, 1956.
- Whorf (1956b). An American Indian model of the universe. In J. B. Carroll (Ed.), *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (pp. 134-159). Cambridge, MA: The MIT Press, 1956.