UCLA
UCLA Electronic Theses and Dissertations

Title
Factors Affecting Learning of Concrete Nouns and Verbs in a Foreign Language

Permalink
https://escholarship.org/uc/item/3c34r8ts

Author
Ludington, Jason Darryl

Publication Date
2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Factors Affecting Learning of Concrete Nouns and Verbs in a Foreign Language

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychology

by

Jason Darryl Ludington

2012

ABSTRACT OF THE DISSERTATION

Factors Affecting Learning of Concrete Nouns and Verbs in a Foreign Language

by

Jason Darryl Ludington

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2012

Professor Patricia M. Greenfield, Chair

First language researchers have proposed dozens of explanations why infants across cultures seem to acquire more nouns than verbs among their earliest words. No such finding has been documented among adult foreign language learners. I wished to determine whether adults have greater difficulty learning verbs than nouns, and if so, why that could be.

To investigate whether adult college students might learn concrete nouns or verbs better, I constructed and measured a list of noun and verb concepts, images, and auditory stimuli. I measured these stimuli on many dimensions with a mind to statistically control these extraneous factors when testing for differences in noun and verb learning (Study 1). In Studies 2 and 3 I trained participants with these words and tested their recognition of targets with multiple-choice tests. I then statistically controlled and measured the effects of measured and manipulated factors to see what confluence of factors affected word learning. While there was generally a noun bias

effect, I found that methods of learning and delay qualified this effect (removing the noun bias at inferential learning, and reversing the word bias at one week), image media quality is of likely help to the learner, and cross-situational learning is greatly helpful. This dissertation concludes with a summary of findings and accounts of them. Although it remained partly unclear why nouns were generally learned better than verbs, Gleitman et al.'s (2006) surface variability hypothesis was taken as the most likely account for the noun bias observed among this sample of young adults.

The dissertation of Jason Darryl Ludington is approved.

James W. Stigler

Catherine M. Sandhofer

John H. Schumann

Patricia M. Greenfield, Committee Chair

University of California, Los Angeles

2012

For my wife.

As I have endured stress from my program, you, also, have endured it.

Let me make it up to you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ACKNOWLEDGMENTS

Thank you, Dr. Patricia M. Greenfield, for your unbounded patience in mentoring me, and for your detailed feedback on all drafts of this and related manuscripts. Only with your skillful mentoring and feedback was I able to advance through the program. Thank you, Kay Lee and Goldie Salimkhan for your amazing artistic contributions to this experiment. Special thanks to Kay, your technical skills and above-and-beyond research assistance were very instrumental in helping me plan and create materials. Thank you to my wife Watanee, you have sacrificed the comfort and security of your family and country to follow me to America to earn this PhD. You have loved me, supported me, and shared in my ups and downs throughout the process. And thank you to my family; you have been a source of encouragement and support throughout my graduate career and throughout life.

VITA

January 18, 1980	Born, Miami, Florida
2003	B.A., Psychology Walla Walla University College Place, Washington
2003-05	Elementary school teacher, Thailand
2005-06	University part-time lecturer, Thailand
2007-09	Teaching Assistant Department of Psychology University of California, Los Angeles
2009	M.A., Psychology Department of Psychology University of California, Los Angeles
2009-11	Teaching Associate Department of Psychology University of California, Los Angeles

PRESENTATIONS

Ludington, J. D., Lee, K., and Greenfield, P. M. (May, 2011). Learning nouns before verbs is more efficient than the reverse order. Poster presented at the meeting of the Association for Psychological Sciences, Washington, D.C.

—, Greenfield, P. M., and Lee, K. (July, 2011). Verbs are remembered better than nouns when inferentially learned. Poster presented at the meeting of the International Association for Cognitive and Educational Psychology, Boston, Massachusetts.

CHAPTER 1: INTRODUCTION AND OVERVIEW

Beginning to learn a new language: Content words

When infants initially speak their mother tongue, they enter a one-word stage followed by a two-word stage of language development (Greenfield, 1976). The kinds of words they first use are not random—they are virtually always frequently used, concrete, content words, in particular mostly nouns and verbs (e.g., "mommy," and "up," as in "pick me up"). It is no mistake that infants speak these kinds of content words first: nouns and verbs convey meaning in and of themselves (unlike function words like "of" or "a," for example), and are among infants' first communicative acts.

More aged learners do not seem all that different. Krashen and Scarcella (1981) theorized adults and children are probably similar with regard to the way that language acquisition proceeds from one-word-at-a-time to more complex usage, if not for different reasons (p.296). From observing my own and others' first attempts to speak a new language, learners begin with single words, then short phrases, and then longer ones. The words they initially speak are normally concrete words (which tend to be more frequent than abstract words). It would be impractical and improbable for learners to begin speaking in full sentences or to begin using infrequent, abstract "higher level" vocabulary words. This is likely to be even truer in unstructured, real-world environments than formal ones. When the primary purpose of speaking is to communicate meaning, first words are likely to be concrete content words, which communicate meaning even as incomplete phrases.

Though there are many important aspects of language development (e.g., learning function words, abstract words, pronunciation, language syntax, morphology), this dissertation

focuses specifically on one of the earliest of developmental steps toward acquiring a new language—learning concrete, content words.

The noun bias debate

Gentner's (1982) seminal research found a noun bias that pervaded six languages she studied: children across most languages and cultures acquire nouns faster than verbs. Verbs are the more difficult of the two word classes for first-language learners to learn. Many theories have been proposed to explain the noun bias in early word learning: location of nouns within utterances (Shady & Gerken, 1999; Au, Dapretto, & Song, 1994; Tardif, 1996; Tardif, Shatz, & Naigles, 1997), the verb argument's requirement of an argument (Greenfield & Alvarez, 1980; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2006; Gentner, 2006; Waxman & Lidz, 2006; Sandhofer & Smith, 2007), inherent verb complexity (Akhtar, Jipson, and Callanan, 2001; Tomasello, 1992), syntactic complexity (Pinker, 1994; Naigles, 1990), cultural emphasis (Gopnik & Choi, 1990, 1995), attention allocation (Kersten, Smith, & Yoshida, 2006), category membership variability (Kersten, Smith, & Yoshida), and natural partitions / relational relativity (Gentner, 1982). As the debate over why the noun bias exists, findings in this dissertation research ought to be useful.

While a great deal of research has focused on what makes words easier to learn in children's first languages, much less research has attended to adults acquiring words in a subsequent language. Does the noun bias apply equally then? This themed question spans this dissertation, although word class was just one of the factors considered.

The importance of word learning predictors

In the study by Gillette, Gleitman, Gleitman, & Lederer (1999) adults with no knowledge of a foreign language whatsoever were better at guessing referents of nouns than of verbs. They

reported on the results of their "Human Simulation Paradigm" in which adults were shown video interactions between mothers and their infants. The videos were silenced, and beeps were inserted in places where the mother uttered "mystery words." The test was to see whether adults might guess the mystery words based on what they could see from the video. Only the most common nouns and verbs were sampled for testing. Gillette et al. showed the beneficial effects of noun knowledge for guessing verbs, of syntactic frame knowledge, and of combined information knowledge, for guessing word meanings. Their study made a powerful statement for the roles of context knowledge and grammar in learning new vocabulary.

The study by Imai et al. (2008) investigated what level of context and grammar support was necessary for 3- and 5-year olds and adults to correctly map words to referents. They found children had less trouble mapping words to objects than to actions. Strangely, adults showed the opposite pattern, mapping words to object referents with 70% accuracy, but with actions they were at 100%. Adults must have approached the vocabulary acquisition task with a different set of assumptions or strategies than were used by the children in that study. In one of their studies Imai et al. demonstrated better performance by Chinese children when verbs were highlighted by video editing to remove the portion in which an object was stationary, but this same method appeared to confuse Japanese and English-speaking children, who performed better when videos segments included the stationary objects. Imai et al.'s study was seminal because, besides showing that we cannot know the learning patterns of adults acquiring a second language based on those of children, Imai et al. also showed different learners bring different strategies to bear on word learning situations. It is the language researcher's responsibility to explore a variety of learning conditions to know what those strategies may be, and which of them work well.

The present dissertation explores and documents the effects of word characteristics (e.g., word class), context characteristics (or characteristics of images, e.g., name agreement), learner characteristics (e.g., sex), and experimental conditions (e.g., ostensive vs. inferential method of learning) that affected young adults' word learning success.

Outline of the present dissertation

Three themes run through this dissertation: the possibility of a divide between noun and verb "learnability" or how easily words can be learned, potential indicators of word learnability, and conditions that lead to successful word learning. Chapter 2 begins with an investigation and collection of characteristics of words, concepts, and images that informed later word learning experiments. Chapter 3 describes a word learning experiment aimed at measuring a hypothesized advantage of learning words by discovering (inferring) their referents rather than having referents simply pointed out. Chapter 4 describes a word learning experiment in which I attempted to elucidate the advantage of learning from two different examples rather than a single example repeated. Chapter 5 is a summary of findings and an account of them.

CHAPTER 2: LEXICAL FEATURES AS POSSIBLE PREDICTORS OF LEARNABILITY (STUDY 1)

One major argument for the greater difficulty to acquire verbs than nouns is their real-world complexity: verbs are ephemeral and hard to point at, but concrete nouns are not (Gentner, 1982, Greenfield & Alvarez, 1980). This view supposes that verbs are harder to acquire because they are harder to identify or parse from sensory information streams. If this view is correct, name agreement of visual stimuli could be a powerful predictor of word "learnability," or likelihood of a word being learned after exposure in a controlled setting. Using line drawing images as a model of real world referents, I measured name agreement, then trained and tested word-to-image referent learning. Name agreement was defined as the proportion of naming responses qualified as target responses. This measure taps the coherence of referent meanings presented through image media, and was used as a proxy for word learnability in the present study. The primary goal of the present study was to measure factors that could affect word learnability. If I were to find that features of words accounted for their accurate identification in images, this would be a major step toward a better understanding of what makes words easier or harder to learn. There was an additional motivation for the measurement of features of nouns and verbs in the present study: I anticipated a learning difference between nouns and verbs, and I hoped to pinpoint why this difference might exist.

I expected to find feature differences between nouns and verbs that would account for differences in learnability. For this reason, in measuring these features, I also contrasted noun and verb values. Features that differ between nouns and verbs become likely candidates for explaining the word class learning disparity in young children.

Investigating factors which may affect recognition and identification performance is commonplace in language research. Research on word decision latency (response time to indicate whether each target is a word or non-word) illustrates this point. Atkinson and Juola (1971), interested in recognition latency, showed that word frequency, concreteness, and word length affected performance. Whaley (1978) demonstrated that richness-of-meaning, letter frequency, and inter-letter probability also affected decision latency. Once these factors are measured and deemed consequential, researchers should try to account for these factors, either by controlling them in their chosen stimuli, or including their measurements in regressions. For example, age of acquisition was controlled when Brysbaert (1996) measured the effect of frequency on naming latency. So important is feature measurement that the goal in some research studies is based primarily on this task. Whaley (1978) expresses the major purpose of her study: ". . . the purpose of the present study was limited to deriving the relative importance of a large set of variables on the prediction of word and non-word classification times" (p. 152). Clearly measuring and determining the predictive value of word factors on outcome measures is valued in the research community. The present study was an endeavor to do just that—to measure a set of variables that could be important for word learning, and to assess their contribution to image name agreement as a proxy for word learnability.

Hypotheses

Prediction 1: Nouns and verbs differ on several features measured

Because the features I measured were likely contributors to word learnability, and because children tend to learn nouns faster than verbs, I reasoned that many of the features I measured would differ between the nouns and verbs in my sample; this would be evidence of their likely contribution to learnability.

Masterson and Druks (1998) found that of their 164 noun and 102 verb images, verbs were rated more visually complex than nouns. Visual complexity might explain naming latency differences between nouns and verbs (Humphreys, Riddoch, and Quinlan, 1988) which might correspond to learnability differences. Concrete nouns are imageable when they can be easily imagined, but concrete verbs may require a bit more finesse in defining because verbs cannot be visualized in isolation, but only by also visualizing agents to enact them. Gleitman et al. (2006) noted that verbs may be harder to learn than nouns because of greater "surface variability in how verbs get realized … within and across languages" (p. 32). Thus I expected ratings of word imageability (how easy it is to conjure up an image of each word) to be higher for nouns than verbs.

Based on past research, I also predicted people would rate noun images as better depicting their intended referents than they would rate verb images (Kauschke & Frankenberg, 2008; Masterson & Druks, 1998). I also predicted that people would offer fewer alternative interpretations for noun than verb images, with target interpretations known. Finally I expected people would name noun images with greater agreement to standard (target) responses than they would with verb images (among children; Kauschke, Lee, & Pae, 2007; indirect evidence in Masterson & Druks, 1998 based on a larger percentage of noun stimuli named at 100% accuracy). Much effort was exerted toward making verb images as identifiable as noun images. However, past findings show verb images to be inherently more complex than noun images (Kauschke & Frankenberg, 2008; Davidoff & Masterson, 1996).

I predicted no difference in concept frequency (the frequency with people encounter the target concepts in their lives) between nouns and verbs. Sandhofer, Smith, & Luo (2000) found from the infant-directed speech of caregivers transcribed and coded in the CHILDES data set that

the most common verbs were used at frequencies nearly equal to those of the most common nouns. My stimuli seemed fairly common, so I did not predict a difference. I also did not predict differences in word familiarity which I assumed to be a product of concept frequency.

Prediction 2: Imageability accounts for name agreement better than its correlates, familiarity and frequency

Gleitman and colleagues (2006) found that word class was not the most important predictor of early word learning; instead they found that "something akin to 'concreteness' rather than lexical class per se, appeared to be the underlying predictor of early lexical acquisition" (p. 27). Their word learning experiment findings, later replicated by Snedeker & Gleitman (2004), showed an overwhelming advantage for nouns. Yet they found that imageability was a better predictor of accurate word naming than whether the word was a noun or verb. Rated imageability is usually correlated with rated familiarity and frequency (e.g., Stadthagen-Gonzolez & Davis, 2006). This might present a challenge for determining which of these imageability-correlates really accounts for outcomes. I tested the effect of imageability, familiarity, and frequency on name agreement, predicting that imageability would account for name agreement better than its correlates. I also explored the other measured features of my stimuli as predictors of name agreement without specific expectations of their effects.

Variables measured

Sources linking variables measured in the present study to learnability were not always available; in such cases a speculative leap was made in predicting their effect on learnability from their known effects on other types of performance (e.g., reading speed). Many of the features measured in this study have never been recognized as predictors of learnability.

Based on the literature certain features of words might be associated with learnability. I measured word familiarity which has been linked to faster reading speed (Brown & Watson, 1987), word imageability which has been linked to lexical decision studies (Balota, Yap, & Cortese, 2006), and concept frequency. Category representation is a variable I conceived to tap the von Restorff effect (1933), the well-established finding that unique items in lists stand out in memory. Category representation was measured as the percentage of targets presented, within a given half of the experiment, (or "list") that were members of each a priori-envisaged category.

I measured name agreement for images of targets in isolation and images of targets in context as the proportion of responses fitting closely with the target response, defined in terms of the coded judgments of several researchers. Name agreement has been shown to affect naming speed (Ellis & Morrison, 1998). I also measured ratings of how well the images conveyed their intended meanings; number of alternative interpretations (raw total number of alternative responses offered by participants who provided this data), for which non-primary responses have been associated with longer response latencies (Székely et al., 2003); ratings of how strange each given noun-verb pair was; and auditory stimulus lengths in terms of utterance time, number of phonemes and number of syllables (for use in later experiments—I did not explore these as predictors of name agreement).

Method

Participants

Thirty participants were recruited from an online recruitment system from a pool of undergraduates in psychology and linguistics courses. One participant's responses were dropped due to that participant not having sufficient English ability (his self-reported ability was below criterion, which was set, a priori, at 8 on a scale of 1 – 10). Not all of the remaining participants

contributed ratings of all stimuli because of inadequate materials preparation, but all 29 remaining participants named images.

Materials

A consent form, biographic data form, and rating sheets were used to collect all hand-written data, and a laptop computer was used to collect naming responses (which were eventually converted to name agreement scores). Name agreement of image stimuli were based on participants' naming responses to images presented using SuperLab (stimulus presentation software) and a Toshiba laptop computer (16:9 LCD display). Following are descriptions of each measurement type.

Biographical data form.

This was used primarily to collect language background information. One question addressed what the participants' first language was. If not English, another question asked participants to rate their language ability in English on a fluency scale from 1-10, where 1 = unable to use any of the language, and 10 = fluent. A third question asked for other languages the participant knew, and how fluent he or she was in each (using this same fluency scale). Age and sex data were also collected.

Images.

Ninety-six (48 noun and 48 verb) black-and-white line drawings images of various everyday items and actions, illustrated in referential isolation, were mostly found on the Internet.[1] These made up a convenience sample based mainly on two criteria: they were concrete

[1] One major source of the images was an online database offering free line drawings of hundreds of objects and actions, along with naming norms, for language researchers by the Center for Research in Language at the University of California, San Diego: http://crl.ucsd.edu/~aszekely/ipnp/. Another resource was simply surfing the Internet using Google's "images" option and filtering to search only black-and-white line drawings. Still other images were hand-drawn by two artistic research assistants: Kay Lee and Goldie Salimkhan.

nouns and observable action verbs, and a decent image of each could be obtained expediently. All objects and actions were of a basic semantic level, not too semantically specific but also not too general. Nouns were mostly animals (e.g., kangaroo) and professions (e.g., doctor), and a few inanimate objects (e.g., refrigerator); verbs were common, familiar actions that could be performed with parts of the human body, such as "to eat" (one exception was "to hatch"). Verb images may be considered images of the present participle. Importantly, verbs were all intransitive verbs, meaning they could be used without specification of a direct object (some could be considered both transitive and intransitive in nature, such as the verb *to write*). While transitive verbs require acting and acted-upon noun arguments, intransitive verbs only require an actor (noun argument in the subject position). By only using intransitive verbs, I could properly present verbs in two-word phrases that made sense without the need of additional information: each phrase described an actor performing an action.

Noun and verb "isolate" images (as I called them, because they were illustrated in referential isolation) conveyed just one elemental concept per image, either a noun or a verb concept. An additional 48 "context" images contained two elemental concepts per image, always an actor performing an action. The names of the elements in these isolate and context image are provided in Appendix A. Some context images and verb isolate images contained a patient (receiver of action) besides an actor, in spite of my trying to select only intransitive verbs; it was hoped the patients of these images would draw less attention than the actor and action would. The noun and verb elements in context images were the same as those in the isolate images. Thus, for example, one isolate image depicted surfing, and another depicted a computer. One of the context images depicted a computer that was surfing. Revision of images was done to maximize name agreement. Toward this end, Adobe Photoshop and Windows Paint were used to

delete background details, crop out any distracting or unnecessary details so that greater attention would be drawn to relevant parts, and add movement marks and lines of motion to verb images to suggest movement interpretations and make these images "come alive," as illustrated in Figure 2.1.

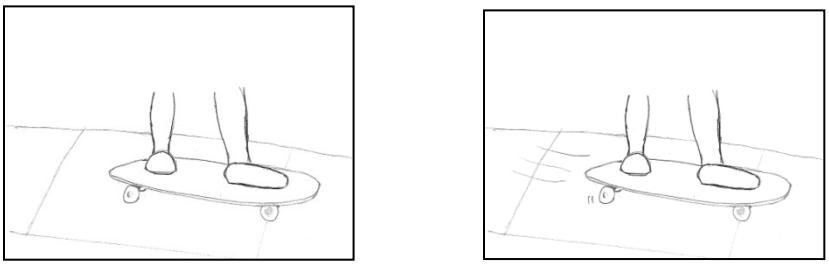

Figure 2.1. "Skateboarding" with (right) and without (left) movement marks. With movement marks the action component becomes more obvious and dramatic. These images are not the actual ones used during experimentation. The source of the original images was lost so to avoid risk of copyright infringement with regard to publication of these illustrations, I drew these images by hand for example purposes.

To ensure that there were no systematic size differences in the noun and verb images, length and width dimensions were measured (using a free application called Pixel Ruler) of all noun and verb images (from top to bottom and left to right of object endpoints, not of image frame endpoints). These dimensions were submitted to an independent samples t tests with an alpha level criterion of p=.05 to mark any differences between noun and verb images. Dimensions were not initially well-matched (noun images were taller than verb images). After shrinking or expanding about 15 noun and verb images, heights and widths were more closely matched. Noun images (457 pixels, SD=88) were still taller than verb images (416 pixels, SD=89), $t(94)=2.27$, SE=18.10, p=.025, but the noun height advantage was offset by a verb

width advantage (450 pixels, SD=92) relative to noun images (421 pixels, SD=104), t(94)=-1.42, SE=20.06, p=.158. Because width and height affect attention almost equally, I added the lengths and widths and compared the summed height-width measurements of noun images to verb images. Noun image dimension sums (877 pixels, SD=100) did not differ from those of verb images (864 pixels, SD=135), t(94)=.516, SE=24.33, p=.607.

Image ratings.

Name agreement. Name agreement was measured as the accuracy of participants' responses to each target, with participant accuracy defined as the average coded judgment of accuracy by six coders.[2] To ensure name agreement was not unduly influenced by participants' knowledge of English, I set the English proficiency criterion to exclude participants reporting less than 8 on the 1 – 10 scale of self-reported English proficiency. Only one participant's data were discarded from the context image naming task for lack of English proficiency (self-reported as 7), and no participants' were discarded from the isolate image naming task.

After 10 of the participants were run, I thought it prudent to make a few small revisions to the instructions and training procedure to improve instructional clarity. At this time I also resized image heights and widths to make noun and verb images more closely matched, as mentioned above. Nine additional participants were run on the improved version of the experiment. To assess whether these minor changes had any qualitative or quantitative effects on name agreement, I performed a 2 (participant group: pilot versus experiment proper, between subjects) x 2 (word class: noun versus verb, within subjects) mixed-subjects ANOVA. Name

[2] Researchers usually calculate a reliability coefficient of coders to be sure they are coding the data the same way when different data are coded by different raters. Here, all data were coded by all six raters, independently. Six judgments should better approximate true values than any one coder's judgment. A mean rating for each word is the best way to represent all six coders' judgments.

agreement in the original, or "pilot" version (M=84%, SD=3.7%) was lower than under the revised version (M=92%, SD=3.5%), but the difference was not significant, F (1, 17) =2.25, MSE=.025, p=.15, and did not change the overall pattern of results, experiment version-by-word class interaction F (1, 17) =.696, MSE=.002, p=.42. Therefore the data were collapsed across participant groups (pilot and experiment proper) to increase power.[3]

Accuracy (agreement of the label with a target label) of participants' naming responses was coded by judges using three values: 0 / .5 / 1, where 0=incorrect, 1=correct, and .5 was reserved for cases that were difficult to classify. A code-book was developed with examples and rationales and a couple general rules: be accepting of morphological variety; if a verb appears as word-class ambiguous (e.g., vacuum could be a verb or noun), take it as correct; accept close synonyms as correct, more distant synonyms as partially correct (.5), but words at the incorrect hierarchical category (super- or sub-categories) as either wrong or partially correct, depending on how distant the relation seems.

Nine research members (eight research assistants and me) shared the response coding burden. Exactly six members coded each participant's responses. Four members coded all responses (i.e., they were members of all sets of coders), while one or two coders were unique to each of three coding teams. Krippendorff's alpha (Hayes & Krippendorff, 2007) was the coefficient used to calculate inter-rater reliability for each of the three overlapping teams of coders because Krippendorff's alpha can be calculated between coders and data sets, does not

---

[3] This decision was validated by later analyses. Comparing correlations between name agreement and other related factors, I observed these correlations strengthen after collapsing across participant groups. The collapsed name agreement values were more strongly correlated with familiarity (r=.558, from .351), imageability (r=.652, from .464), and frequency (r=.373, from ns). However the correlation between goodness of representation and name agreement grew weaker (r=.600, from .668). It seems highly unlikely that weak or nonexistent correlations would grow stronger after more data are gathered, compared to when fewer data are gathered. It is thus more likely the case that name agreement as a construct is really correlated with these other measures, and that collapsing data across experimental procedures improved these correlations by improving measurement accuracy without compromising the validity of the measurements, justifying the decision to collapse across these minor procedural differences.

require each code to be used by each coder at least once, and is robust to missing values. To avoid violation of non-independent judgments made when a research member coded target responses repeatedly for different respondents, one alpha was calculated for a set of six coders for each respondent's data. Thus a total of 29 alphas (19 for isolate images, 10 for context images) were found based on 29 participants who contributed data, and these alphas were averaged. Table 2.1 presents the alpha averages for the three coding teams. Krippendorff (2004) has suggested an alpha standard as .800 or higher when reliability is crucial, and alphas of .667 as useful for providing tentative conclusions, but that no magical cutoff number exists. In the present case, because coding agreements are not crucial to the overarching goals of this dissertation, the obtained alphas were deemed acceptable. The alpha average in the isolate image naming condition was .71 (for 19 respondents), and the average alpha in the context image naming condition was .66 (for 10 respondents). I averaged codes across the six coders and across all respondents to obtain a measure of name agreement of each target. Isolate and context image name agreement averages for nouns and verbs are provided in Appendix C.

Table 2.1

Krippendorf's Alphas[a] *for Each Set of Six Coders' Judgments*

Image type	Coder #	N	Krippendorff's alpha
Isolate	#1-4, 8, 9	10	0.78
Isolate	#1-6	9	0.64
Context	#1-4, 7, 8	10	0.66

[a]Codes (0 / .5 / 1) were considered to be on an interval scale.

Goodness of depiction. Ratings of "goodness of depiction" or how well each isolate image depicted its intended target were collected with the question "How well does the image represent the concept?" on the following scale: "1=not at all, 2=not well, 3=somewhat, 4=pretty

well, 5=very well." Goodness rating averages for noun and verb images are provided in Appendix B.

Number of alternative interpretations offered. Participants were also asked to offer alternative interpretations of each isolate image if they thought of any. These responses were collected right after each item was rated for its goodness of depiction. I measured this variable as the sum total of all alternative interpretations offered by all participants for each image. In cases where multiple participants offered the same alternative interpretations, the number of alternative interpretations was counted as number of offered responses, and not the number of interpretations; in other words I counted the quantity of participants, as well as the quantity of responses of each participant, in calculating each sum. Values ranged from 1 (e.g., computer) to 17 (to snort). Averages for noun and verbs images are provided in Appendix B.

Concept ratings.

A stapled set of sheets was provided to participants to collect the following word measurements. *Concept frequency* ratings were taken to assess "How often have you encountered these concepts—either directly or in images—over the course of your life?" on the following scale: "1=never, 2=rarely (once every 2 years or less), 3=frequently (once every 6 months or less), 4=quite frequently (once every month), 5=extremely frequently: once every week at least." Participants rated *word familiarity* using the question, "how familiar is each item to you" by entering a number from 1-7, where 1=completely not, and 7=completely. The same participants that rated familiarity also were asked to rate *word imageability* on "How easy is it to generate a mental image of each item" using this same scale.

Another form assessed *strangeness of stimulus pair*, asking "How strange are these two-word concepts?" (1= "completely natural," and 7= "completely strange, I would never expect

these two concepts to be together"). Some of the "stimulus pair" or context images showed an actor performing an action that is within the normal range of activities commonly performed by such actors (e.g., cat sleeping), whereas other images showed actors performing actions that are uncharacteristic or unrealistic given the actors (e.g., hippo knitting). It was thought that the strangeness of each concept pair might affect how well participants could infer or remember the meanings of word-referent associations.

I measured *concept representation* of each English word stimulus as the percent of concepts fitting in that word's given "category." I conceived of three noun categories—humans, all other animals, and non-living things—and four verb categories—performed by arm (or hand or finger), performed by leg (or foot or toe), performed by face (including by eyes, nose, ears, mouth, or tongue), and performed in some way that could not be classified in these mentioned categories. These categories were the most evident to me given the stimuli. I classified the word typing as an action performed by arm/hand/finger, and classified the word refrigerator as a non-living object, as examples.[4]

Nonsense words.

Auditory stimuli were created, 96 nonsense words in total, 48 randomly assigned as noun labels, and 48 as verb labels. I decided not to use an existing language because real languages contain a mix of familiar and unfamiliar phonemes and phonemic structures, variables that I wanted controlled. Instead I created a mix of one- and two-syllable words to simulate words of a real language.

Nonsense words were created using a pool of 17 consonant phonemes and 7 vowel phonemes following a CVC (one syllable) or CVCVC (two syllables) structure to simulate words

[4] The categorization of the refrigerator was made in spite of artistic personification of this and other non-living objects by illustration of non-living objects using human characteristics, such as eyes, mouth, etc.

of a real language. Table 2.2 lists the consonants and their position rules. The set of consonants for word-initial and word-final positions was made partially overlapping, as is often the case in real languages. In many languages, some phonemes never or rarely occur in initial or final locations (e.g., "ng__ in English, and "__r" in Thai). Nonsense words were largely adopted from Vitevitch and Luce (1999) or adapted from the same phonemes they used to make their nonsense words, which were constructed from highly common in English phonemes. Highly familiar phonemes are more perceivable than unfamiliar phonemes (Appleman & Mayzner, 1981), though they may not be any easier to remember. Balanced numbers of one- and two-syllable words were created and assigned as nouns and verbs, 36 one-syllable and 12 two-syllable words for each word class.

Table 2.2

Phonemes Used to Construct Nonsense Word Stimuli

Location in Syllable	Phonemes
Beginning	D, F, G, H, J, K, L, N, P, R, S, Sh, T, Th, W, Y, Z
Middle	Ai, Ee, Eh, Ir, O, Oo, Uh
Ending	B, Ch, D, F, G, H, Jsh, K, L, M, N, P, S, T, Th, V, Z

Nonsense words were spoken by a native English-speaking Caucasian adult male (me) and recorded using Audacity 1.3.12 (Beta) (a free sound recording software) which was also used to edit and measure utterance lengths of all auditory stimuli. Sounds were edited to include a 100 milliseconds onset delay so that they would not be sounded simultaneously with image onset (to avoid distraction or reduced attention to either sense modality, either sight or hearing, when presented together). Individual words were recorded for presentation with isolate images, and two-word phrases were recorded for presentation with context images. I spoke and recorded

phrases with normal sentential intonations (as a continuous utterance, not staccato words) to maintain the ecology of stimuli as complete phrases.

Word utterance lengths were measured to the nearest hundredth of a second. These measurements were submitted to an independent samples t-test to determine whether nonsense words assigned to one word class or the other could have coincidentally differed in utterance lengths. Nouns (M=.91, SD=.19) did not take significantly more time to utter than verbs did (M=.96, SD=.22), independent $t(94)=-1.05$, SE=.042, $p=.30$. Words were randomly assigned to concepts with the aid of random.org (a free online randomization engine). I called this assignment of meanings to nonsense words "Language A." After this I performed a second random assignment in which I randomly swapped all labels between nouns and verbs to create a second nonsense word language, "Language B."

Procedures

Upon entering the experiment room, participants signed a consent form and completed a biographical data sheet. Participants were then randomly assigned to either identify the elements in isolate images or context images, and tested individually. Next participants sat down at the computer and read directions that were presented on computer. The directions read as follows:

> You will see a series of line drawings each depicting a *single* [or *pair of*, for those assigned to name context images] English word[s]. At the same time, you will hear a [pair of] non-English word[s] that means the same thing as the English word[s]. Please write what you believe each non-English word means IN ENGLISH. If you aren't sure, guess. DO NOT WRITE THE WORDS YOU HEAR. First you will complete 8 practice trials. The research assistant will coach you through this portion.

The nonsense words were played at the onset of each image only to simulate the conditions of future studies in which I anticipated presenting auditory sounds with images.

Participants completed a few practice trials (4 nouns, 4 verbs; this meant 8 practice trials for those assigned to name isolate images, but only 4 practice trials for those assigned to name context images). During practice trials, the experimenter provided correctional feedback to participants' incorrect responses. If the target was a verb, the corrective feedback was stated more or less as follows: "Actually the intended response was [noun] or a [noun] / [verb]-s or [verb]-ing." This was done to highlight the noun or verb nature of images, and to train participants to identify images in a way that would be less ambiguous for coding.

In the isolate image condition 48 objects and 48 actions were presented, totaling 96 images (in addition to practice trials). In the context image condition 48 images were presented, each containing an actor performing an action. Images were presented to individual participants while a pre-recorded auditory stimulus presented nonsense-word labels of the objects at the onset of each image display on the screen. Images remained on the screen until participants typed a response and pressed "Enter".

After completing the naming phase of the study, word and image ratings were collected. Finally participants were debriefed and given course credit. Participation was always completed within one hour.

Results and Discussion

The means of measures (familiarity, frequency, imageability, isolate image name agreement, context image name agreement, goodness of representation, number of alternative interpretations offered) were calculated for nouns and for verbs. Table 2.3, below, shows this information as well as the scales of measurement, and results of independent samples t tests done

to compare nouns to verbs on each factor. Results are discussed with reference to the predictions laid out at the beginning of the study.

Table 2.3

Sample Ns, Means, Standard Deviations, t- and p-values of Seven Stimulus Factors

			Nouns		Verbs			
Factor	Scale	N	Mean	SD	Mean	SD	t	p
Familiarity	1 to 7	26	6.45	0.84	6.49	0.62	-0.45	0.66
Frequency	1 to 5	20	3.87	0.43	3.43	0.55	-2.18	0.03
Imageability	1 to 7	26	6.67	0.48	6.68	0.35	-0.09	0.93
Goodness	1 to 5	20	4.91	0.09	4.75	0.22	2.81	0.01
Alternatives	raw #	20	3.77	1.98	5.6	3.2	-3.37	0.00
Isolate Naming	0 to 1	19	0.90	0.13	0.86	0.11	1.94	0.06
Context Naming	0 to 1	10	0.90	0.07	0.78	0.15	3.32	0.00

Prediction 1

I predicted nouns more than verbs would be rated as more imageable, and rated as better depicted with fewer alternative interpretations offered—all features that suggest greater learnability.

Word imageability

Imageability data were submitted from 26 participants. For the 96 words rated, the 48 nouns were rated no more imageable (M=6.67, SD=.47) than the verbs (M=6.68, SD=.37), independent samples t(94)=-.093, SE=.087, p=.93, contrary to prediction. By this measure, my attempt to develop a set of nouns and verbs for subsequent experiments that were equally imageable was successful.

Goodness of depiction

Ratings of how well each image depicted its intended target (target labels were provided below each image) were collected from 20 participants. Consistent with prediction, noun images

(M=4.92, SD=.15) were rated as better depictions of their target concepts than verb images were (M=4.77, SD=.33), independent samples t(94)=2.81, SE=.052, p=.06.

Number of alternative interpretations offered

Right after rating how well images depicted intended targets, the same 20 participants also responded with alternative interpretations of these illustrated concepts by answering the question "Can this image represent other concept(s) besides what it is said to represent? Please indicate." As predicted, fewer alternative interpretations were offered for noun images (M=3.77, SD=1.98) than of the verb images (M=5.60, SD=3.20), independent samples t(94)=-3.37, SE=.543, p=.001.

Name agreement

Isolate images.

Nineteen participants contributed name agreement responses to isolate images. Consistent with prediction and with many naming studies (e.g., DeBleser & Kauschke, 2003; Kauschke and Frankenberg, 2007; Davidoff & Masterson, 1995), nouns were better identified than verbs, as measured by average judged name agreement with target responses. An independent samples t test demonstrated that nouns isolate images (M=.91, SD=.07) may have been identified better than verb isolate images (M=.87, SD=.13), t(94)=1.94, SE=.021, p=.06, though this difference did not quite reach significance. Figure 2.2 portrays means and standard errors of measurement of noun and verb isolate image name agreement. Appendix C provides the name agreement values for all isolate images.

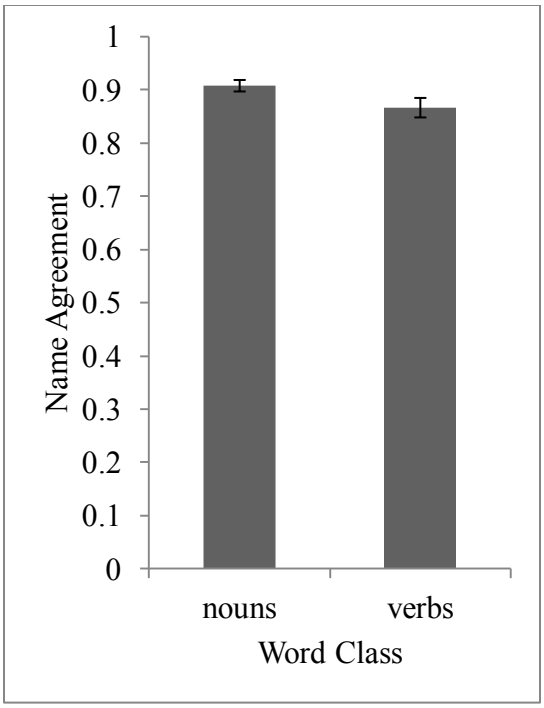

Figure 2.2. Isolate image name agreement of nouns and verbs. Error bars represent standard errors of measure.

Context images.

Eleven participants participated in the name agreement measurement task for context images. One participant's data were excluded due to insufficient English proficiency (7 on a scale of 1 – 10; 8 or higher was chosen as criterion, a priori). The 10 remaining participants' responses were analyzed. Noun and verb name agreement means are displayed in Figure 2.3. In these 48 context images, the noun target aspects (M=.91, SD=.13) were more accurately identified than the verb target aspects (M=.79, SD=.22), independent samples t(94)=3.32, SE=.037, p=.001, in accord with prediction and with the literature (e.g., DeBleser & Kauschke, 2003; Kauschke and Frankenberg, 2008; Davidoff & Masterson, 1996). Name agreement values for these 48 context images is provided in Appendix C.

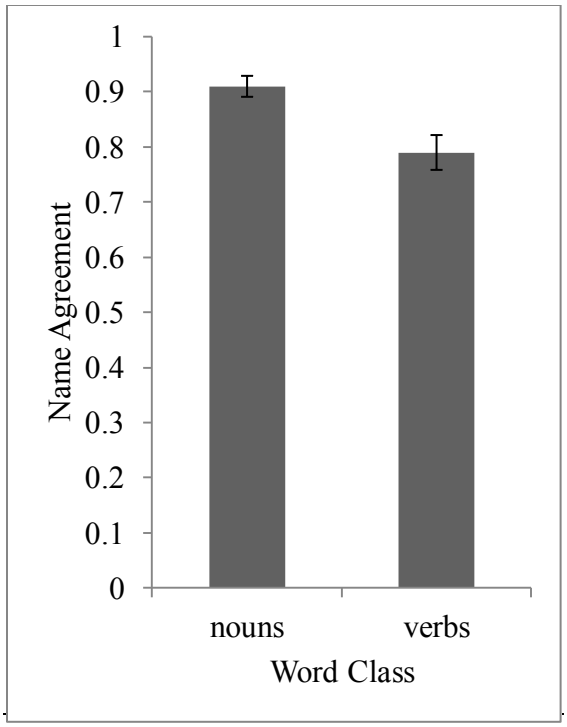

Figure 2.3. Context image name agreement of nouns and verbs. Error bars represent standard errors of measure.

Next I analyzed whether the noun in each context image was systematically more nameable than the verb in order to see about the possibility that the observed differences were due to a chance few poor images. Given the low sample size (N=10), and the nature of the coding scale that was used (0, 0.5, and 1, averaged over 6 coders) a Wilcox signed ranks test for related samples, which assumes data are on an ordinal scale, was used to account for the size and direction differences in name agreement between nouns and verbs in each image to determine if the name agreement differences were systematically one way more than the other, across images. This test showed that on average the noun aspect was more nameable than the verb aspect in context images, more than would be expected by a chance few poor drawings, Wilcox signed ranks test of 48 images, $Z=-3.6$, $p<.001$. Thus the noun naming effect found with isolate images

extended to context images, and this advantage was not likely the result of a chance few poor line drawings.

Concept Frequency

In the introduction I suggested that some features would not differ between nouns and verbs. I predicted no word class difference in frequency ratings. Twenty participants completed concept frequency ratings. Contrary to prediction, the verbs (M=3.87 (on a scale of 1 – 5), SD=1.03) were rated as more frequent than the nouns (M=3.43, SD=.94), independent samples t-test, t(19)=-2.18, SE=.201, p=.03. This finding is not inconsistent with word frequency characteristics in English, however; very common verbs tend to occur a little more frequently than very common nouns (Sandhofer, Smith, & Luo, 2000).

Word Familiarity

I also predicted no word class difference in word familiarity ratings. The same twenty-six participants as above also rated word familiarity. Nouns (M=6.45 (on a scale of 1 – 7), SD=.54) were not rated different than verbs (M=6.49, SD=.53), independent samples t(94)=-.45, SE=.108, p=.66, in line with prediction.

Prediction 2

One major purpose of the present study was to better understand the features of stimuli that make their image identification more accurate. Attributing explanatory value to a variable is difficult when related variables share similar relationships with the criterion variable, however. Past studies have found familiarity, imageability, and frequency share correlations with one another (Bird, Franklin, & Howard, 2001). Indeed these factors were correlated in the present data set (all rs > .64). But I predicted imageability would account for name agreement better than frequency and familiarity (the imageability-correlates). Using ordinary least squares regression I

tested this prediction, first with isolate image name agreement as the criterion, and second with context image name agreement as the criterion.

Isolate images.

I first tested imageability and its related predictors familiarity and frequency in three individual predictor models to see they were significantly related to isolate image name agreement. Familiarity was related to the criterion, $r=.49$, $p<.001$. So, too, were frequency, $r=.30$, $p=.003$ and imageability, $r=.59$, $p<.001$, but imageability was most highly correlated with name agreement, in support of my hypothesis that there was a stronger correlation between imageability and criterion than between familiarity or frequency and criterion. Next I entered imageability, familiarity, and frequency into a single model of isolate image name agreement to see if some portion of imageability explained name agreement beyond what imageability shared with familiarity and frequency in predicting name agreement. The model as a whole was significant, $p<.001$, and imageability was the only significant predictor of accurate name agreement of targets in isolate images in this model: imageability, $p=.002$, familiarity $p=.93$, frequency $p=.38$, strengthening my position that imageability is the most fundamental of these predictors of name agreement. From the one-factor model, imageability was shown to explain a sizeable portion of isolate image name agreement variance, adjusted r^2 of naming agreement = .34, F change $(1,94)=50.43$, $SE=.086$, $p<.001$. Therefore among this sample of 96 words, imageability seemed to explain about 34% of the variation in participants' isolate image name agreement. Figure 2.4 illustrates this, showing that as words were rated as more imageable, participants were more likely to name these targets in images correctly.

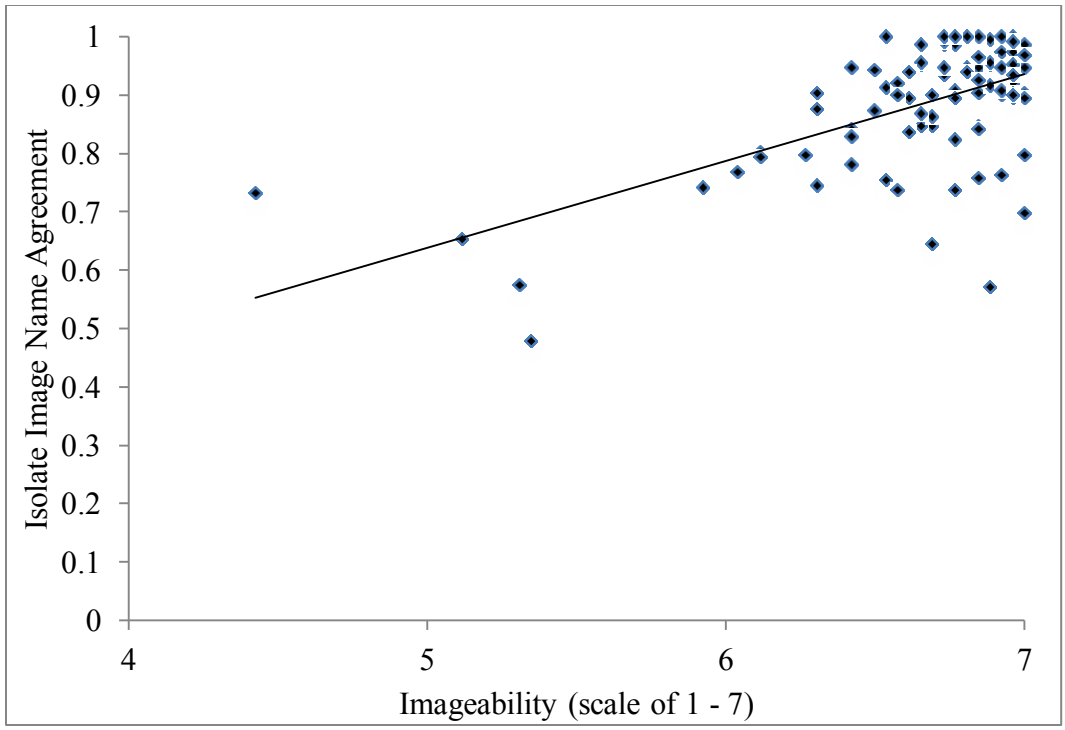

Figure 2.4. The relationship between word imageability and isolate image name agreement. The abscissa are shown plotted on a shortened axis to magnify this portion of the 1 – 7 scale (the minimum imageability rating was 4.42 for "hedgehog"), while the ordinates are shown plotted along their full scale.

Context images.

Using the same method as above, familiarity, frequency, and imageability were initially entered individually as predictors of context image name agreement in three 1-predictor models. Familiarity ($r=.29$, $p=.004$) and imageability ($r=.36$, $p<.001$) were significant, but frequency was not ($r=.07$, $p=.52$, all two-tailed). Next I placed familiarity and imageability in the same model as predictors of context image name agreement. Imageability was significant, $p=.02$, but familiarity was not, $p=.49$. Thus imageability accounted for context image name agreement better than its correlates, as predicted, and above and beyond familiarity. The adjusted r^2 for imageability in its

own model was .12; thus about 12% of the variance in context image name agreement could be explained by word imageability. This relationship is shown in Figure 2.5.

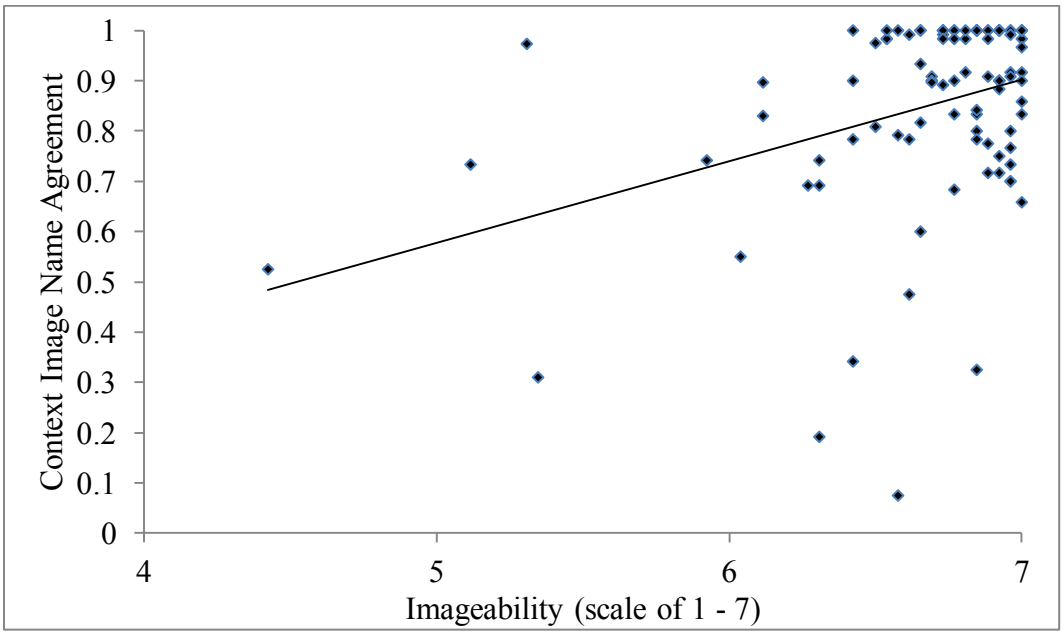

Figure 2.5. The relationship between imageability and context image name agreement. Again the abscissae are shown plotted on a shortened axis to magnify this portion of the scale (the minimum imageability rating was 4.42 for "hedgehog"), while the ordinates are shown plotted along the full scale.

I also tested the relationships between name agreement and all of the other feature measurements taken. The results of this exploration are reported in Table 2.4, below. Category representation was poorly correlated with name agreement indices (both ps>.05), and therefore a poor candidate for explaining learnability. However goodness of depiction and number of alternative interpretations were both strongly related to both name agreement indices. These two measures were related, r = -.69, not surprisingly; because they were measured together (as the procedure went) as two ways to capture how well images depicted referents as intended. Thus in

addition to the effect of imageability, two measure of image goodness also functioned to explain a portion of name agreement.

Table 2.4

Correlations Between Two Name Agreement Indices and Three Features of Stimuli

	Name Agreement Indices					
	Isolate Images			Context Images		
Variable Name	r	adj r^2	p	r	adj r^2	p
Category representation	.13	.01	.20	.18	.02	.09
Goodness of depiction	.52	.26	<.001	.54	.28	<.001
Alternative interpretations[b]	-.39	.15	<.001	-.53	.27[a]	<.001

[a]Adjusted r^2 values can be translated directly as a percent of name agreement variance accounted for by variables named at left.

[b]Note that this table is not a single model of name agreement, and therefore the variance explained by one factor (e.g., goodness of depiction) is unlikely to be independent of the variance explained by another factor (e.g., alternative interpretations).

Next I asked which better accounted for name agreement, either goodness of depiction or alternative interpretations. When both predictors were regressed on isolate image name agreement, the model was significant with goodness of depiction as a significant predictor in the model, $p<.001$, but not number of alternative interpretations, $p=.59$. In its own model, goodness of depiction accounted for 26% of isolate image name agreement. When both predictors were regressed on context image name agreement, they were both predictive; goodness was again reliable, $r=.33$, $p=.005$, but number of alternative interpretations also independently predictive in the model, $r=-.29$, $p=.01$. The model's adjusted $r^2 = .32$, meaning together these two predictors accounted for 32% of the variance in context image name agreement.

Finally out of curiosity I tested and found isolate and context name agreement values were correlated at $r = .45$, $r^2 = .20$, $p < .05$. The existence of a correlation between name agreements for images related by their concept suggests the concept itself explains about 20% of the variation in name agreement.

Conclusions

The noun images in my sample were rated as better depicted, participants offered fewer alternative interpretations, and nouns were identified more accurately in both isolate and context images than verbs were—in spite of my research associates and I going to great lengths to attempt to equate verb with noun name agreement. These differences suggest nouns may be better learned from these images than verbs. In some other ways the nouns and verbs did not differ: no differences were found in imageability, familiarity, or frequency ratings. However, only imageability, goodness of depiction, and number of alternative interpretations were significant predictors of name agreement. Therefore based on this study, if a word class difference in learnability were to be found among this sample of words, it would most likely owe to word class differences in goodness of depiction and number of alternative interpretations, because these were also significant predictors of name agreement.

Numerous image revisions were made to try to make verbs as identifiable as nouns (but the verbs were still less accurately named than the nouns). Over the course of these revisions, my research team learned that verbs usually are not easily recognized in images unless they are "animated" with certain, often-used tricks employed by cartoonists and artists. We used graphic motion cues including lines to indicate from where movements originated, and marks near moving parts, to introduce dynamism into still images. While these symbols are not found in the ecology of the real world, I perceived they facilitated interpretation of motion from still images. I

did not measure name agreement rates before and after verb image doctoring, so improved name agreement by employing these symbolic devices for now remains an impression. If these and other symbolic devices really did elevate name agreement, an educational application based on this possible effect could be proposed.[5]

The measurement of features of nouns and verbs was useful to subsequent investigations of word learnability which I describe in the next few chapters. These measurements may also allow other researchers to equate or otherwise statistically control stimuli on these factors which are typically extraneous to purposes of investigation.

[5] If images can be used to conveying word meanings without need to reference any other languages, this could be very useful. The use of images rather than words could allow materials to be presented with almost any population with few needed changes. Also, because images are highly memorable (Lutz & Lutz, 1978), image media are probably a highly useful means of teaching foreign vocabulary.

CHAPTER 3: MODELING LEARNABILITY: TESTING 23 PREDICTORS OF NONSENSE WORD RECOGNITION (STUDY 2)

My experiences teaching English as a second language in universities in Thailand impressed upon me how valued and vital second language teaching and learning is across the globe. This impression led me to the literature on second language acquisition among adults and children. Across this literature I have found that a "noun bias" among early learners is a well accepted phenomenon, but among adults such a word class advantage is virtually unheard of. I decided to experiment upon adults to test for a possible presence of this word class advantage. I used image media as a model for natural and instructional forms of word learning.

The noun bias debate, spanning three decades, aims to address why children initially learn more nouns than verbs. But does this "nouns-earlier" phenomenon extend to adults learning a second or subsequent language? Some researchers have argued that nouns are initially learned better because verbs, as relation words, are too ambiguous until the learner has some knowledge of relate-able parts. With the present experiment I aimed to shed light on this hypothesis.

In this study I adopted the fast-mapping paradigm (Carey & Bartlett, 1978) to test questions relating to word class learning differences; participants saw images and heard words labeling targets in those images. As is often done in fast-mapping studies where labels are learned for objects or actions, participants were given forced-choice tests wherein words learned prior were presented again, and target selection was from among several choices (e.g., Markman & Wachtel, 1988; Yu & Smith, 2007; Alishahi, Fazly, & Stevenson, 2008; Imai et al., 2008; Vlach, Sandhofer, & Kornell, 2008; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; O'Hanlon & Roberson, 2007).

I tested word knowledge by word recognition performance, a form of word knowledge, rather than a recall test because I thought cued recall tests would be too insensitive to the kind of under-developed knowledge forms I expected participants would have after learning many novel words from few meaningful examples within limited time periods in the present paradigm (i.e., I wanted to avoid floor effects). In this dissertation I do not make much distinction between word recognition and word learning. This view is defensible: in cognitive psychology recognition and recall are frequently used as analogs of one another to assess subjects' states of knowledge. Recognition and recall have long been conceptualized as two features of the same construct—memory. Testing recognition rather than free recall is much like using a stethoscope rather than an EKG monitor to measure someone's pulse—though both measure the same construct, they do so by measuring different biophysical phenomena.

Purposes

Testing for a noun bias among adults learning foreign vocabulary

Children typically learn nouns faster than verbs in learning their first language. Does this learning bias hold for adults learning a subsequent language? Why?

Differentiating between two compelling hypotheses

I addressed several questions in the present study that may have implications for word learning in natural and formal environments. One of these questions is the role of prior noun knowledge in verb learning. Two propositions were put forth in Greenfield & Alvarez (1980). According to what I have called the "parts before relations" hypothesis, verbs (relations) make more sense in the presence of relatable nouns (parts). This hypothesis is difficult to test on first language learners because one cannot teach or test a verb without using at least one noun argument to convey a verb's meaning (e.g., a young child cannot learn or choose 'jumping'

without someone or something enacting jumping). This hypothesis may account for the higher proportion of nouns to verbs in young children's vocabularies, but may or may not explain adult second language vocabulary development. Verb concepts learned in the second language might have translate-able equivalents in learners' native language(s), allowing verbs to make sense independently of nouns. Thus the parts before relations hypothesis can be tested on second language learners to see whether learning a noun helps verb learning in a special way, or if noun knowledge aids verb learning just as much as verb knowledge aids noun learning.

A second proposition put forth in Greenfield and Alvarez (1980) is what I have called the "referential ambiguity" hypothesis. In their study, Greenfield and Alvarez found that as the number of unknown words and meanings in a context decreased, learning increased. Verb learning situations always arise with actors, and sometimes with patients and tools of enactment; these potential learning situations may arise with too much referential ambiguity to allow learners to effectively map verbs to their meanings. Prior learning of these contextual arguments can aid learning of a verb in these situations by disambiguating which one of multiple possible referents maps correctly to the verb word. Noun knowledge can function to reduce ambiguity of verbal referents, but verbs may not serve as well in the same way to aid noun learning because nouns are typically less ambiguous to begin with. However the verbs in the present study were also fairly unambiguous because graphical cues were added; teaching meanings with these images could challenge the parts before relations hypothesis. Also, by using verb images with known name agreement indexes, ambiguity could be controlled and tested to highlight the importance of disambiguation in word learning. According to the referential ambiguity hypothesis, the roadblock to learning label meanings is the lack of clarity with which labels refer

to perceivable referents; in this account the primary "on" switch for learning is reducing ambiguity for referents.

To set these hypotheses upon one another, I manipulated prior knowledge to assess its role in vocabulary learning. To manipulate prior knowledge of nouns versus verbs, a noun or a verb image was ostensively presented—meaning it was presented by itself—and a novel word was uttered therewith. After its presentation, this word became "knowledge." This allowed seeing how this (now assumed) knowledge affected learning of an additional word. I sought to know whether noun knowledge helped in learning a verb more than verb knowledge helped in learning a noun. I predicted no difference.

This is not what Greenfield and Alvarez (1980) found; however, their verb images drew attention to several object components besides verbs, which made their verb referents, in some cases, highly ambiguous. My verb images were created to draw attention to verb components using graphical cues, thereby disambiguating verb referents. In the present study knowledge of another word should not disambiguate verbs much further because they have already been disambiguated graphically. Thus I predicted learning my verbs would not depend on prior noun knowledge any more than learning nouns would depend on prior verb knowledge. This prediction comes directly from the referential ambiguity hypothesis—when ambiguity is reduced (by any means—in this case by graphic cues), conditions favor learning. This prediction differs from that predicted by the parts before relations hypothesis, under which learning nouns before learning verbs is more effective than vice versa.

Comparing the effectiveness of two methods of learning

Another purpose of the present study was to contrast two learning methods to determine the more effective one. A direct and commonsense approach to teaching anything is to just teach

it rather than to "teach around it." However there may be redeeming value in an indirect approach. The indirect technique used in the present study was to expose learners to an unknown element in a controlled context, allowing learners to discover an indirectly-presented meaning. The process of discovering meaning might lead to a deeper kind of processing that facilitates word memory.

I manipulated method of learning at two levels. I call the direct method "ostensive" learning and the indirect method "inferential" learning. In the ostensive conditions, each label (a noun or a verb) and its referent was learned from individual presentation of them together. Inferential conditions were created so that participants would have to infer the meaning of one element presented in a context image. I predicted greater performance under inferential learning conditions because I presumed meaning discovery through inference would be a deeper mental process than ostensive (associative) learning.

Partway through data collection I recognized the need for an additional manipulation of ostensive conditions. Ostensive conditions were initially presented with a redundant labeling using a context image right after each isolate noun-verb image pair's presentation, but inferential targets were not redundantly labeled. To provide orthogonal manipulation of the number of occurrences of each target within its trial apart from ostensive conditions, I ran an additional experimental condition with additional participants in which words were presented ostensively without redundancy. Procedural differences between the initially-begun experiment and the added condition were minor enough to warrant inclusion of both conditions under the same experimental name, thus adding a variable to the study, "number of occurrences." This added variable was tested to address the question, Does the number of examples and occurrences of a

target have an effect on word learning? I predicted greater performance when targets were presented with two occurrences than when they were presented with only one.

Ostensive and inferential learning conditions are illustrated in Figure 3.1, below. Ostensive conditions always included isolated examples, called "isolate images," of each target; for some participants these isolate images were followed by a redundant example called a "context image" which contained both prior-named isolate image elements. In inferential conditions, a single isolate image and a context image were shown, in that order, which allowed learners to recognized the redundant element and thus infer the meaning of one of the two target elements from their context.

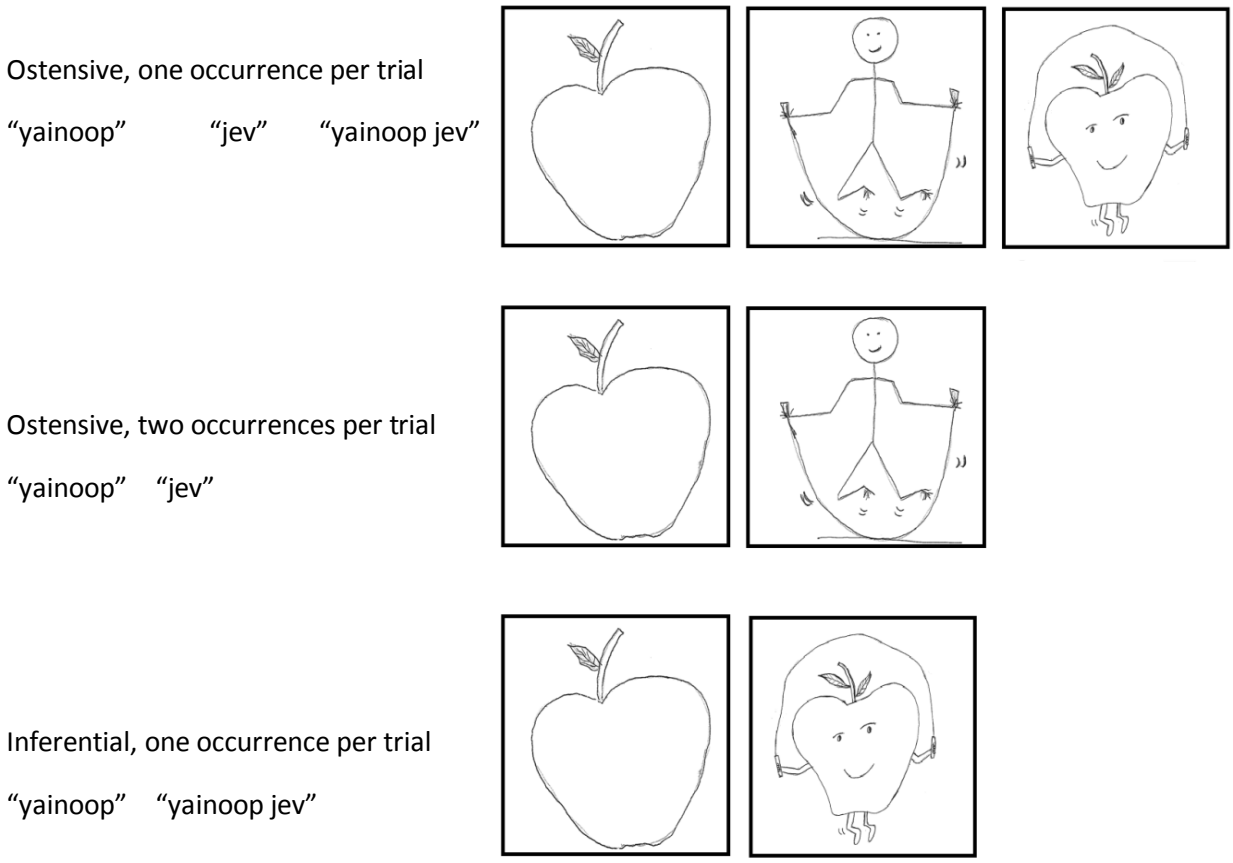

Figure 3.1. Flow diagrams illustrating ostensive (top and middle panels) and inferential learning (bottom panels) of the verb “jev,” meaning “to jump rope”. The isolate images of the apple and of jumping rope were obtained with permission from Arlene Arthur (a close friend) to publish them, and I drew the image of the apple jumping rope.

Notice in Figure 3.1 that in the inferential condition (shown in the lower panel in the figure), a learner should gather that “jev” means “to jump rope.” The learner can infer this based on the principle of mutual exclusivity (Markman & Wachtel, 1988), an application of the following four-step logic: 1) “yainoop” is known, based on the first image, to refer to the apple; 2) “yainoop” cannot refer to jumping rope because it refers to the apple; 3) “jev” should not refer

to apple because "yainoop" refers to apple; therefore 4) "jev" must refer to jumping rope because that is an otherwise unnamed referent in the last image.

The possibly uneven effect of method of learning across word classes.

An important purpose of this study was to investigate whether inferential learning was better suited for learning verbs than nouns. The parts before relations hypothesis stipulates better verb inferential learning because nouns knowledge aids learning verbs but verb knowledge is not as useful for learning nouns. The referential ambiguity hypothesis predicts no interaction between the method of learning and word class; whatever word class learning difference may exist in ostensive conditions should replicate at inferential conditions. Verbs were made unambiguous by graphical means; therefore if reduction in referential ambiguity is sufficient to explain learning, nouns and verbs should be learned about equally well inferentially because verbs were already rendered as unambiguous with graphical motion cues. The presence or absence of an interaction between method of learning and word class, therefore, would provide support for one but not the other of these two hypotheses.

The possibly uneven effect of method of learning across delays.

Another important purpose of this study was to address a question regarding the effect of delay on recognition. I measured learning at two learning delays—five minutes and one week. By measuring twice, I could test for possible differences in retention under ostensive and inferential learning conditions. Most fast-mapping studies test participants within a matter of seconds or minutes, but at such short delays it is questionable whether fast-mapped words are truly learned or may be just temporarily held in working memory in the auditory loop. A few other studies have demonstrated that fast-mapped learning does remain measureable after longer delays, anywhere from five minutes (Horst & Samuelson, 2008), one or two days (Jaswal &

Markman, 2003, experiment 1; Woodward, Markman, & Fitzsimmons, 1994), to one week or longer (Carey & Bartlett, 1978; Vespoor & Lowie, 2003), but very few studies have manipulated learning methods at multiple delays.

Knight (1994) is one notable exception: Knight found that ostensive learning (by glossing with a dictionary) aided learning more than inferential learning (guessing meanings from a reading, followed by confirming or disconfirming by glossing) on an immediate test and on a one-week delayed test. But Knight's inferential condition was not purely inferential; it was really inferential and ostensive learning together, and it took longer than either method by itself.

I wanted to test the effect of delay to see if pure inferential learning modulated the effect of delay on memory, so I manipulated test delay at two levels, five minutes and one week. After five minutes I predicted participants would recognize more words learned ostensively than inferentially, as Knight found, because inferential learning necessarily involves some uncertainty of mapping words to their referents, while ostensive learning involves greater certainty of associations. But after one week I expected better recognition of words learned inferentially than ostensively due to slower forgetting of inferentially learned words. Inferential learning seems to require more processing than ostensive recognition because inferential learning requires guessing and confirmatory testing for every inference. More involved or effortful processing is typically associated with longer term retrieve-ability (Bjork, 1994). As a deeper level of processing, I predicted the memory trace of inferentially learned words to be better detected than the more shallow ostensive form of learning at one week.

Other potential predictors of word learning

This chapter also addresses the question of what, besides word class, method of learning, and delay, are some other predictors of successful word recognition. It was this question that

motivated the measurement of many features of stimuli in the previous chapter and the measurement of characteristics of participants and conditions in the present chapter. I used an exploratory approach to test and model the effects of a total of 23 factors that might matter to word learning. A summary of all questions addressed in Study 2 is given below.

1. Is there a noun bias among adults learning foreign vocabulary?
2. Does the method of learning matter? Does it affect the rate of forgetting?
3. Is one method of learning better for verbs than nouns? Is one method better for nouns than verbs? Which hypothesis (parts before relations vs. reduction in referential ambiguity) do the results better support?
4. What are some other predictors of word learning? These variables emerged from the prior study and included ratings of familiarity, frequency, imageability, goodness of image representation, number of alternative interpretations offered, strangeness of word pairs, category representation, isolate and context image name agreements, and utterance lengths of words.

Method

Participants

Ninety participants from the University of California, Los Angeles were recruited with an online recruitment site as used in the previous study. Some participants were dropped due to participants' failure to return for the second part of the experiment (5) or for experimenter failure to present all materials (1). The mean age of the remaining 84 participants was 20.8 years, SD=4.19 years. More females (62; 73.8% of sample) than males (22) participated. Most participants were at least partially able to use a second language (only 2 did not report any second language ability). The average number of languages (including English) reported at any proficiency on a 1–10 scale was 3.5. Participants reported their proficiencies in all languages

including English on a 1-10 scale. The average language proficiency sum across participantsand across all languages besides English, based on the aforementioned scale, was 7.8 (SD=4.5).

Design

Twenty-three predictors were tested in total, about half of them continuous, and half categorical. Some were experimentally manipulated, and some were not. These factors are described in detail in the subsequent sections of this chapter. Most of the categorical factors were manipulated within participants. Word class (nouns versus verbs), method of learning (ostensive versus inferential, as defined in previous chapter), and number of word lists learned prior (none versus one) were all manipulated within subjects. Experiment languages, bothcreated from the same stock of nonsense words (Language A vs. Language B), and learning schedule (schedule 1 versus 2; these are defined below) were manipulated between subjects. Test delay was manipulated within-subjects for 56 of the participants, but was fixed at the five-minute delay for 28 of the participants. Number of occurrences per trial was fixed at only one occurrence or two occurrences among these 28 and 56 participants, respectively.[6] Word order (order with which isolate images were presented: noun-verb versus verb-noun) was counterbalanced within participants and was another within-subjects categorical variable. Word order and number of occurrences were analyzed separately because both were nested within the ostensive level of the method of learning. Participants' sex and English-as-first-language status (English as most proficient language versus not) were also measured.

[6] After 56 participants were run, I recognized a missing condition and ran 28 participants through a supplemental condition—the ostensive conditions among these participants contained only one target occurrence per trial. This provided an orthogonal manipulation of method of learning apart from number of occurrences, between subjects. These 28 participants were tested only at five minutes, and not at one week, because of data collection time concerns.

There were a number of continuous variables tested as well: participants' age, self-rated English proficiency (scale of 1–10), and other language proficiency sum (the sum of all self-reported, self-rated proficiencies in languages other than English) were measured and tested in the present study. Additionally the following stimulus measurements, garnered and defined in Study 1, were tested: category representation, word imageability, word familiarity, concept frequency, goodness of depiction, number of alternative interpretations, word-pair strangeness, name agreement in isolate images as well as context images, and utterance lengths measured as number of phonemes, syllables, and time of utterance (in seconds and hundredths of seconds).

The dependent variable was the correctness of each target word selection (i.e., recognition) made, measured on a binary scale as correct or incorrect. A 25% likelihood of target selection marked chance performance, as the recognition task was to select the correct target from among four targets. This variable, measured for each word learned, was nested within each participant (i.e., each participant was measured multiple times). Thus it was the word unit, nested within the participant unit, which was analyzed.

Materials

Variables.

The variables explored in the present study as potential predictors of word recognition are listed in Table 3.1 below. Some were defined and measured in the previous study (ratings of familiarity, frequency, imageability, goodness of image representation, number of alternative interpretations offered, strangeness of word pairs, category representation, isolate and context image name agreements, and utterance lengths of words). Given the number of factors considered, I relate details—coding and analytical procedures—in context with results for a simpler organization of information.

Table 3.1

All 23 Factors Explored in Study 2

	Item Factors	Participant Factors
Images concepts	Familiarity	Age
	Frequency	Sex
	Imageability	English-as-1st-lang status
	Category representation	English proficiency
	Name agree, isolate	Language proficiency sum
	Name agree, context	
	Goodness of image	Condition Factors
	Alternative interpretations	Method of learning
	Strangeness of context	Word class
	Utterance-length	Word order
		Delay until test
		Experiment language
		Lists learned prior
		Occurrences/trial
		Learning schedule

Nonsense words.

Ninety-six words (with utterance lengths measured in Study 1) and their 48 phrases were presented with images.

Languages.

Languages A and B were counterbalanced between subjects. Language A was formed by randomly assigning 96 nonsense words to all 96 targets, with one- and two-syllable words counterbalanced between nouns and verbs. Language B was then formed by randomly re-assigning noun nonsense word labels given in Language A to verb targets, and verb nonsense word labels given in Language A to noun targets.

Syntax.

Words in phrases were ordered in a noun-verb typology (as English uses), such as "[A] doctor [is] smoking." Thus although words were initially presented in either a noun-verb or a verb-noun order, contextual utterances were always uttered in noun-verb order.

Images.

The images used in this study were the same as those measured and described in Chapter 2. The noun and verb means on some important characteristics of these images are provided in Appendix B.

Learning.

The learning program.

A Toshiba laptop computer (screen size: 19 inches diagonally) was used to present words, sounds, and images using Superlab 4.0. Details of how this program presented stimuli to participants follow.

Events. A learning event was composed of an image and auditory stimulus (nonsense word) presented at the same time. The nonsense word's onset was purposely recorded with about 100 milliseconds of silence at the beginning of each sound clip so that words were not sounded simultaneously with the onsets of images. Events advanced over time at a rate of one event every three seconds.

Trials. Each learning trial was composed of a set of events, either two or three, presented sequentially, so each trial lasted either 6 or 9 seconds. Trials were presented consecutively as a continuous progression of images throughout each segment.

Target occurrences within trials. Participants learned words in ostensive segments and inferential segments within subjects. Among ostensive segments, targets were presented either

once or twice in each learning trial; this variable, number of target occurrences per trial, was manipulated between subjects. Among participants who viewed targets with two occurrences per trial, each target's second occurrence marked the trial's end and thus would have allowed participants to parse trials and group words according to trials as pairs. However among participants who saw targets presented only once per trial, the presentation format was such that image pairs may not have been identified by participants as pairs, per se. That is, where no target occurred more than once per trial, trial end-points were less apparent, and participants presumably did not parse trials very well, which would have made grouping targets as pairs unlikely. This design cost was outweighed by its design advantage—enhanced control and comparability between methods of learning, and between levels of target occurrences. One-occurrence trials and inferential trials both contained two events; number of events controlled, so any recognition difference would owe to an effect of method of learning. One-occurrence and two-occurrence ostensive trials both contained ostensive presentations of the same isolate images; any difference to be found between these conditions would owe to the presence or absence of a context image in trials because the same isolate images (and characteristics of those images) were seen between these conditions.

Blocks. A block was composed of 8 trials. The same block was shown 6 times repeatedly with no breaks between block repetitions in each segment. Each time a block was shown, its trials occurred in a different randomized order.

Segments. A segment was composed of a block repeated six times. Each segment presentation lasted from five to seven minutes, and was entirely made up of either ostensive trials or inferential trials, but never both. Each segment was preceded by two practice trials demonstrating the pattern of that segment's image progression. Between segments, participants

engaged in a distractor task, attempting a Sudoku puzzle, for 30 seconds to reduce learning and attention fatigue, pro-, and retroactive interference from other blocks.

Two inferential learning segments and one ostensive learning segment were viewed at each learning session. The reason for twice the number of inferential to ostensive learning segments was to obtain equal numbers of data points from the two methods of learning. Twice as much ostensive data were collected per ostensive trial as inferential data were collected per inferential trial because each ostensive trial presented two words which could be tested, while each inferential trial only allowed inference of one word, so only that one could be tested from a given inferential learning trial. The two inferential learning segments were always presented consecutively to reduce number of instructional changes between segments. The ordering of segments was ostensive, inferential, inferential at one of two learning sessions, and inferential, inferential, ostensive at the other. Segment orders and order of segment orders were both counterbalanced between learning sessions.

Lists. Ninety-six words were evenly divided into two lists. The reason for two lists instead of one was to enable manipulation of the test-delay variable within subjects, and to reduce the number of targets per lists to make each learning session more manageable for participants. Each list was divided into three learning segments. Every participant learned both lists and thus saw six segments.

Learning schedules.

Participation occurred in two parts, both involving learning and testing. The testing effect (improved recall for already-tested items) was not an issue because no item was tested twice. Among 56 participants, half of the words were tested after a five minute delay, and the other half of the words were tested after one week. For the remaining 28 participants, the test delay was

fixed at five minutes. To limit proactive and output interference for either list, a learning schedule was utilized in which 28 of the participants learned the first list at their first of two appointments and the second list at their second appointment one week later. The remaining 56 participants learned both lists within a single appointment. The two learning schedules are illustrated in Table 3.2. Schedule 1 might lead to more output interference due to learning both lists nearer in time, but less proactive interference due to the length of time intervening between learning both lists, while Schedule 2 might lead to more proactive interference but little output interference. Use of both schedules helped to even these effects out.

Table 3.2

Learning Schedules

Schedule	Learn	Delay [a]	Test	Learn	Delay	Test
1	list 1	1 week	list 1	list 2	5 minutes	list 2
2	list 1	5 minutes	list 1	list 2	1 week[b]	list 2

[a]The bold, zigzag line segment demarcates what occurred during the first (to the left of the line segments) and second appointments (to the right of the line segments), separated in time by one week.

[b]Half of the 56 individuals assigned to this learning schedule actually experienced a second five minute delay (not a one week delay as the diagram shows).

Instructions.

Participants read segment-specific instructions relating to each segment's learning condition. In the ostensive learning block the instructions read, "In this section, slides are ordered into TRIPLETS presented back-to-back: 1st – word 1 is spoken (you will see an illustration of it), 2nd – word 2 is spoken (and illustration), 3rd – a phrase is spoken containing those words again (and illustration). Both words are equally important." Among the sample of

participants who were presented with ostensive conditions containing only one target occurrence per trial, directions read "In this section, slides are presented back-to-back." For the inferential learning blocks, instructions were as follows: "In this section, slides are ordered into PAIRS presented back-to-back: 1st – a word is spoken (you will see an illustration of it), 2nd – a phrase is spoken containing that word AND another word (and an illustration of them). Both words are equally important." After the above sets of instructions tailored to conditions were presented, some general instructions followed: "You do not need to respond. Just learn what the words mean. Later, you will be tested! Practice 2 triplets [inferential condition: "pairs"] first. The experimenter will guide you during this practice." After two training trials were shown, the following text appeared on the screen: "Can you tell the experimenter in your own words what you will be doing in this block?" Feedback was provided to clarify the instructions as necessary.

Testing.

The test was conducted using Superlab 4.0. English words, arranged vertically as numbered options, were presented in the center of the screen in Times New Roman 18-point font. At each item onset, a sound file presented a nonsense word from the most recent learning session. Two tests were given to assess learning of the two word lists. There were 96 test items testing all learned words.

Foils were chosen from among the learned stimuli so that all choices would be equally familiar. One of the foils was always a meaning that co-occurred with the target in its context image during learning. Each English word was offered four times at test, once as target, and three other times among foil options.

The participant's task was to indicate which English word was referred to by the spoken word (a forced choice paradigm) using four number keys ([1], [2], [3], and [4]) to designate

choices. Two practice test items were given, the choices and targets derived from the learning practice trials. No breaks were given during testing. Participants' responses were scored by the presentation software as correct or incorrect.

Procedure

Participants were tested in a departmental lab space by any of seven research assistants or me. Participants who arrived at the research site in a timely manner were randomly assigned to a learning schedule. Those who arrived more than ten minutes late *and* indicated they had another engagement at the end of the hour were assigned to the schedule that would allow them to complete the first part of their participation within that hour (fewer than 10% of the sample). The effect of learning schedule on word recognition was assessed.

Upon entering the testing room, participants completed a consent form and filled in a language and biographical data form that asked for their sex, age, primary language, proficiency in English, other languages spoken, and proficiencies in those languages. Next participants were seated at a computer and the experiment was started.

Participants completed two learning portions and two testing portions of the experiment according to the schedules shown in Table 3.2. When participants experienced the five-minute delay, they were told to play Tetris for five minutes before being tested. When they experienced the one-week delay, they left for that day and returned seven (minimum six, maximum eight) days later. Afterwards participants were debriefed and credited.

Results and Discussion

I collected recognition responses for all words presented to all participants. Due to experiment programming errors, some test items were presented with options in which the target was absent. This occurred on 4 items under Language A, and on 10 items under Language B for

two-thirds of the participants (these errors on the test were corrected prior to running the final one third of participants); data for these problematic test trials was removed before analyses. An alpha criterion of .05 was used to determine significance for all assessments.

Logistic regression allowed me to take advantage of known qualitative and quantitative differences amongst stimuli, participants, and conditions to test their predictive value. Twenty-three factors were tested, and models were developed to describe their effects under the present paradigm.

Analytic strategy

Modeling effects of 23 variables was tricky with a total sample size of only 84. To use a step-wise regression (which allows a statistical program to determine a model that fits the data best given a set of variables) would not have been desirable given the sample's size—the risk of false discovery would have made interpretation difficult. Exploratory analyses (i.e., data mining) require very large sample sizes to counteract inflated false-discovery rates. I also did not go to the other extreme—if I were to test only a few, very particular word learning models, error rates would not have posed much threat to interpretation, but I would miss the chance to explore the predictive characteristics of a number of variables that I had measured. Instead I used my background knowledge of word learning to set forth on an educated exploration of the predictive worth of factors and interactions that I thought might be of interest. I decided which factors to include or exclude at certain steps in model development based on significance values. Given the number of factors analyzed and tests performed, logistic analyses presented here must be taken with grains of salt because error rates were inflated by the number of analyses performed.

As a first step I explored all individual predictors. Second, I tested interactions that would be interesting, or that I had formed specific predictions regarding. Third, I modeled the

categorical predictors, found significant in step one, together. Then I did the same for the continuous predictors. Fourth, I combined the significant predictors of the categorical and continuous models with the significant interactions I had explored in the second step to check that all effects found thus far survived controlling for all other effects found thus far. This led to formation of an over-sized model of word learning. Finally, as this model was too large for the number of participants measured,[7] I divided this model into two smaller models describing recognition measured at the two testing delays. I used an alpha criterion of .05 to make decisions to retain or discard factors during all factor testing and model development.

Results of testing each factor in its own model are reported in Table 3.3 with Wald χ^2 values and p values to indicate reliability of improvement in individual model predictions over null hypothesis predictions based only on the grand mean. Effect sizes are not reported until later when more developed models are presented because simple one-factor models tend to over- or under-estimate effect sizes when multiple effects are involved.

Also shown are the interactions I tested between some of these factors and delay. For each test of an interaction, the two factors and their interaction factor were modeled together. Only the p-values of the interaction components of these models are cited in the table. Note that interactions with delay were not tested exhaustively—some factor interactions with delay were

---

[7] Just how many factors may be included in a model? A fairly common rule of thumb regarding sample sizes needed for regression analysis is N > 10k when there are k predictors; this would allow up to 8 factors in my study with 84 participants. Green (1991) more conservatively proposed a rule of thumb where N > 50 + 8k (this would translate to only 4 factors in my study); and Hosmer and Lemeshow (2000) advised N = 20k in logistic regression (also translating to 4 factors in my study). But Vittinghoff and McCulloch (2006) found that such rules of thumb may be too restrictive. In the present case, the sample size was 84, so in light of Vittinghoff and McCulloch's research, and given the number of observations per participant, I felt comfortable limiting the number of factors in models to eight or fewer. In exploratory research this ratio of N : k can be smaller, but as it gets smaller, generalizations beyond the sample become riskier (Berger, 2003).

not tested either because they were unlikely, were correlated and thus partially accounted in another tested variable, or because they were not manipulated orthogonally with delay.

Table 3.3

All 23 Factors Tested with Individual Models

	source	Main effects Wald χ^2	p	Means (SD), or % of cases	Interaction w/ delay p
Participant	Age	0.01	0.94	20.22 (1.76)	n.t.
Factors	Sex	1.47	0.22	72.6% female	0.89
	Eng. as 1st lang	0.19	0.66	79.8% English 1st	0.05
	Eng. proficiency	0.12	0.73	9.64 (1.01)	n.t.
	Lang prof. sum	0.05	0.82	17.95 (4.42)	0.66
Condition	Delay	306.68	0	67.1% 5-min	n.t.
Factors	Class	0.2	0.66	50.2% nouns	0.007
	Method of learning	27.48	0	67.1% ostensive	0.01
	Order	5.18	0.02	50.0% n-v order	0.73
	Experiment lang	0.02	0.88	50.2% Language A	<.001
	Lists learned prior	63.32	1	65.7% no prior list	<.001
	Learning schedule	4.41	0.04	33.3% schedule 1	0.09
	Occurrences/trial	4.46	0.03	44.4% 1 occur/trial	0.57
Item Factor	Category represent.	0.01	0.92	0.19 (.09)	0.24
Conceptual	Familiarity	5.53	0.02	6.47 (.51)	0.63
	Imageability	7.56	0.01	6.68 (.40)	0.35
	Frequency	4.08	0.04	3.65 (.99)	0.93
Visual	Goodness depiction	1.89	0.17	4.85 (.26)	0.91
	Alternatives	7.36	0.01	4.65 (2.76)	0.01
	Strangeness	0	1	4.52 (1.87)	n.t.
	Name Agree-Isolate	2.85	0.09	0.88 (.14)	0.14
	Name Agree-Context	3.34	0.07	0.84 (.19)	0.86
Auditory	Utterance length	0.62	0.43	0.93 (.20)	0.06

[a]n.t. means not tested

Main effects.

Categorical variables.

Most of the categorical factors were significant in their own models. These were significant: delay, method of learning, word order, number of lists seen prior, number of target occurrences per trial, and learning schedule. Effect sizes are provided in more developed models.

Next these significant categorical predictors were combined into a single model. Only four factors remained significant beyond one another. Order and method of learning were treated in separate models (with all the other components the same) because order was nested within, and not crossed with, the ostensive level of method of learning. In a three-factor model, delay, method of learning, and number of occurrences were all highly reliable factors (all $p<.001$) in a highly significant model, Wald $\chi^2(3)=344.42$, $p<.001$. Order, when modeled with delay and number of occurrences, was also a significant component, $p=.046$.

Continuous variables.

None of the characteristics of participants were predictive of recognition success (age, English-status, English proficiency, total language proficiency sum). Familiarity, imageability, and frequency were all significant predictors individually (all $p<.05$), however the results of Study 1 informed that these were correlated. Regressing moderately or highly correlated items together reveals the value of each beyond others, but common variation is removed, often resulting in failure to meet significance criteria. To get around this, I selected one spokes-factor of these correlates to include in later, more developed models. To determine which among related factors should be their spokes-factor, I tested these three in a model together to see if one could account for recognition above the others. None could, but I found that imageability ($p=.11$) was closer to significance than its correlates familiarity ($p=.51$) or frequency ($p=.48$). I also tried

a composite of these three correlates, but the composite was less reliably predictive than imageability, so I elected imageability as the spokes-factor among its correlates. This finding is reminiscent of the finding in the previous study—that imageability predicted name agreement better than its correlates.

Goodness of depiction and number of alternative interpretations were both significant predictors, but these, too, were highly correlated and should not progress to more developed models together. To decide which factor to elect for later model testing, I tested these two together and found that number of alternatives remained significant (p=.02) beyond goodness, which was not significant in their model. Thus I selected number of alternatives as the spokes-factor for these two.

Neither of the name agreement indices were significant in their own models, but they both closely approached significance. Knowing they were correlated, I modeled them together, but of course neither of them was significant beyond the other. Because of their theoretical importance I selected one of these to progress to the next phase of model development in spite of poor individual reliability. I chose context image name agreement, even though it was not the more significant of the two, because context images were more viewed.[8] Next I modeled all significant, continuous spokes-factors together. Context image name agreement lost its trend toward significance, now p=.42, so I eliminated it from the model. Imageability (p=.02) and number of alternatives (p=.03) were both significant beyond one another, Wald $\chi^2(2)=12.44$, p=.002.

[8] Two-thirds of participants viewed 100% of context images, and the other one-third of participants saw 66% of the context images. All participants viewed only 66% of isolate images; this amounts to isolate images viewed by fewer participants.

Interactions.

In testing an interaction between two predictors, STATA (statistics computation software) automatically includes their individual factors in the model. Here I do not report on the significance of individual predictors because I have done so earlier. I also do not report Wald values of these interactions because these were based on the whole model itself rather than on any specific factor. I tested the prediction that the effect of method of learning would vary by word class. This interaction predictor was significant (p=.04), supporting the parts before relations hypothesis.

Additionally I explored many factors that might or might not interact with delay because this might reveal predictor values more or less robust to delay. I found that many factors interacted with delay: word class, method of learning, experimental language, number of lists learned prior, number of alternative interpretations, utterance length, and whether English was one's first language. These interactions suggest uneven predictive values at the two test delays.

Combining these individual interactions in a single model, the interactions between delay and utterance length (p=.14) and between delay and number of alternative interpretations (p=.72) lost significance. Pulling these out, the developed "interaction model" was reliable, Wald $\chi^2(12)=379.65$, $p<.001$, with all remaining interaction factors significant: delay-by-English-status (p=.003), delay-by-word class (p=.009), delay-by-method of learning (p=.01), delay-by-experimental language (p<.001), delay-by-number of lists learned prior (p<.001), method of learning-by-word class (p=.03). However, this model included more terms (12) than should be typically used with samples of this size (N=84).

Modeling categorical, continuous, and interaction predictors together.

Next I combined the significant predictors in the models of the categorical predictors, continuous predictors, and interaction predictors. All factors were significant except one: the interaction between delay and method of learning became insignificant (p=.06). The effects of order (p=.04) and number of alternatives (p=.04) were also of questionable reliability. However I included these three in the model shown in Table 3.4 because they were on either side of the significance threshold. Fifteen predictors were used in this model (some are insignificant, included because they were involved in interaction effects), which is more than is typically allowed; therefore this model may not generalize beyond this sample. The effects of method of learning, word class, delay, and their three interaction effects (delay x method of learning, delay x word class, method of learning x word class) are illustrated in Figure 3.2, below the table, as odds ratios (odds are relative to the given reference group) on an odds metric. Appendix D offers a short introduction or refresher on odds ratios for readers less familiar with this metric of analysis. Figure 3.3 shows the model-based effects of the same factors on probability of recognition.

Table 3.4

*Over-sized Model Including All Significant Predictors (*Wald $\chi^2(15)=399.59, p<.001$*)*

			Measured values	
Predictor	p	odds ratio	min	max
Delay	<.001	0.28	5 min	1 week
Method of learning	<.001	0.49	ostens	infer
Target occurrences per trial	.002	0.74	1	2
Imageability	.01	1.17	4.42	7
Alternative interpretations	.04	0.98	1	17
Word class	.04	0.86	noun	verb
Word order	.04	1.17	n-v	v-n
English-as-1st-language status	.76	0.94	yes	no
Experimental language	.22	0.82	A	B
Lists learned prior	.09	1.14	none	one
Delay X Eng 1st language status	.004	0.62	-	-
Delay X word class	.01	1.32	-	-
Delay X method of learning	.06	1.24	-	-
Delay X experimental language	<.001	1.80	-	-
Delay X lists learned prior	.001	0.61	-	-
Word class X method of learning	.04	1.24	-	-

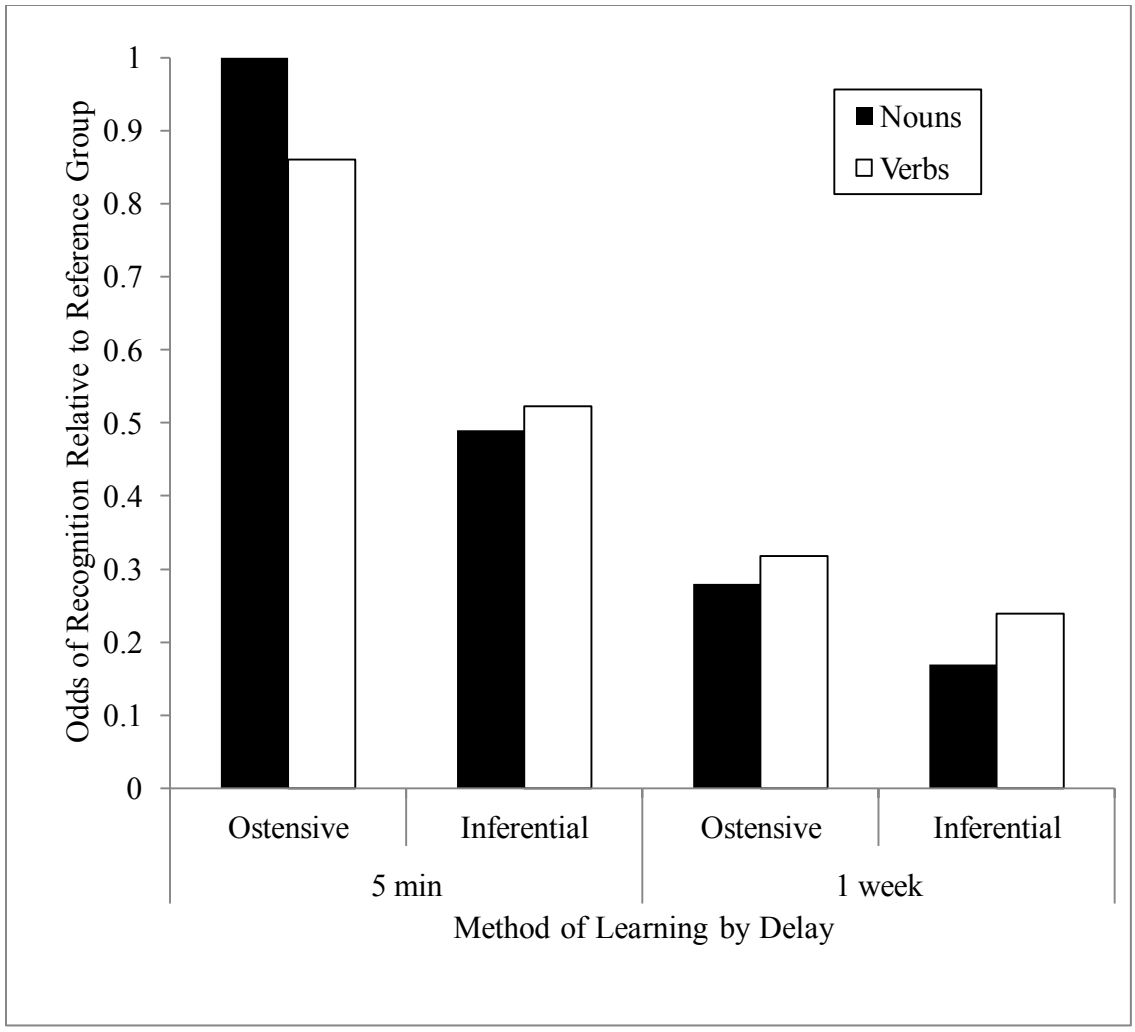

Figure 3.2. The odds ratios of method of learning, word class, delay, and all three of their interaction effects on the odds (based on the over-sized model) of recognition. All other model factors are controlled in this illustration. Odds are illustrated relative to ostensively learned nouns at five minutes (which represents the reference level of all three factors).

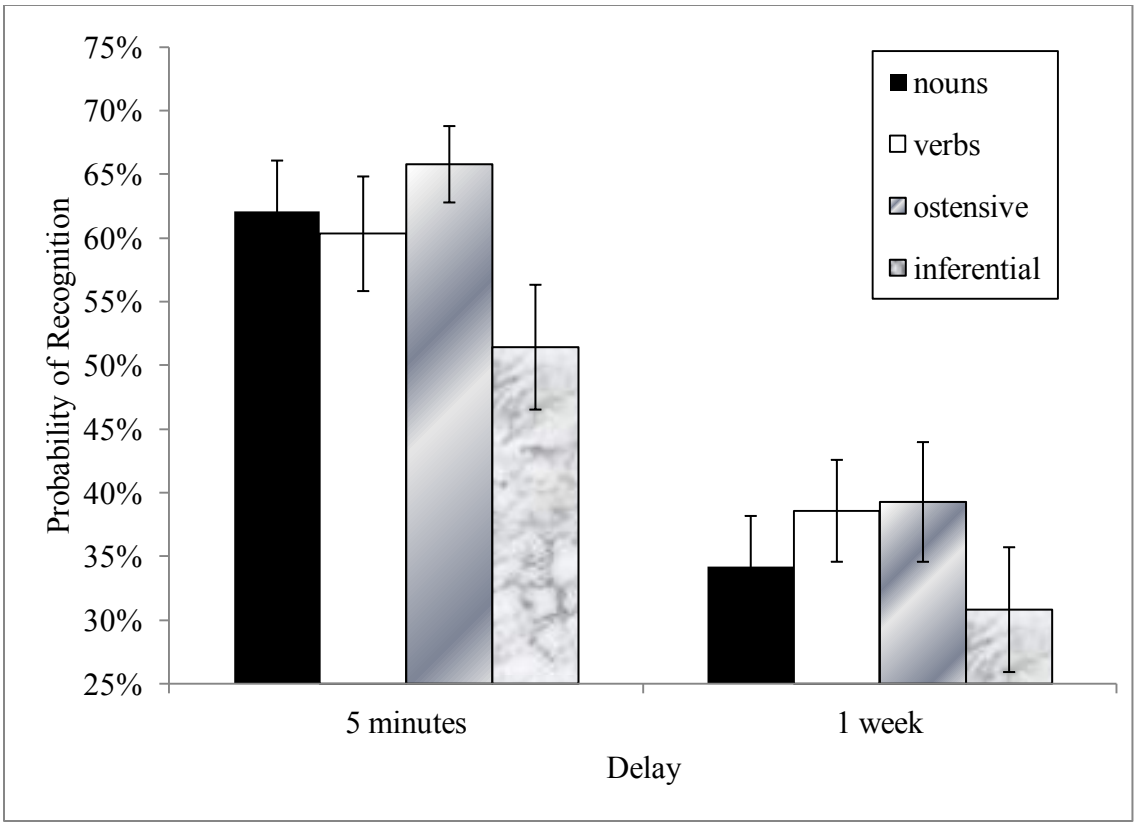

Figure 3.3. The over-sized model-specified effects of method of learning, word class, and delay on the probability of recognition. Values shown include contributions from all other factors of the over-sized model. The ordinate axis is shortened to only 25 – 75% for closer inspection of effects, and because chance performance was at 25%. Error bars represent 95% confidence intervals.

Five-minute model.

As this model was over-sized, I decided to divide it into two smaller, more manageable models by separating the data by delay condition and re-running analyses on smaller resulting models. Dividing the data along the delay dimension allowed the removal of all interaction factors that included delay, greatly reducing the total number of model predictors.

Dividing the data by delay, there were only seven factors from over-sized model to test: method of learning, word class, word order, number of occurrences per trial, number of

alternative interpretations, imageability, and the interaction between method of learning and word class (word class was not significant in the over-sized model, but was a significant part of this model). Testing these on the data collected at five minutes, word order lost its significance (p=.12), as did context image name agreement (p=.88), so I removed these from the model. I compared values of other factors before and after removing word order and name agreement to be sure their presence or absence from the model did not alter other modeled factors. Their removal had no effect on any other modeled factor. Number of alternative interpretations also lost significance when tested only with the five minute data (p=.16). However I decided against removing this factor from the model because with its removal, three other model factors increased in reliability: word class, imageability, and the interaction between word class and method of learning. Therefore by including number of alternative interpretations in the model, the reliability and effect size of other modeled effects were adjusted conservatively. I called the resulting model the "five minute model," which is shown in Table 3.5 below with significance values and odds ratios. Note that odds ratios for method of learning and word class are not straight-forwardly interpretable because of the interaction component included in the model.[9]

[9] The odds ratios in the table are ratios of odds at alternative values over odds at reference values. Odds ratios of factors involved in interactions are the odds ratios of a factor when its interacting factor is held at reference. In Table 9, the odds ratio listed for method of learning is when word class is held at reference level (nouns), and the listed odds ratio of word class is when method of learning is ostensive. (The odds at the alternate level of both interacting variables are found by multiplying both factors' odds ratios and the interaction odds ratio together.)

Table 3.5

Five Minute Model of Word Recognition (Wald $\chi^2(5)=67.00$, $p<.001$)

			Measured values	
Predictor	p	odds ratio	min	max
No. of occurs / trial	.003	0.74	1	2
Imageability	.004	1.25	4.42	7.00
Method of learning	<.001	0.45	osten	infer
Word class	.01	0.82	noun	verb
Word class X meth-of-learn	.01	1.39	-	-
No. of alternative interpret.	.16	0.98	1	17

When all other things were controlled, two occurrences per trial was associated with decreased odds of recognition by a factor of .74 relative to one occurrence per trial, oddly. (I inspected actual numbers of correct and incorrect responses at both occurrence levels and confirmed the direction of this effect.) Perhaps redundancy within trials was a learning turnoff. Imageability had a strong positive effect of increasing the odds of recognition by a factor of 1.25 for every one unit increase in imageability. This factor's effect on the probability of recognition at five minutes is illustrated in Figure 3.4.

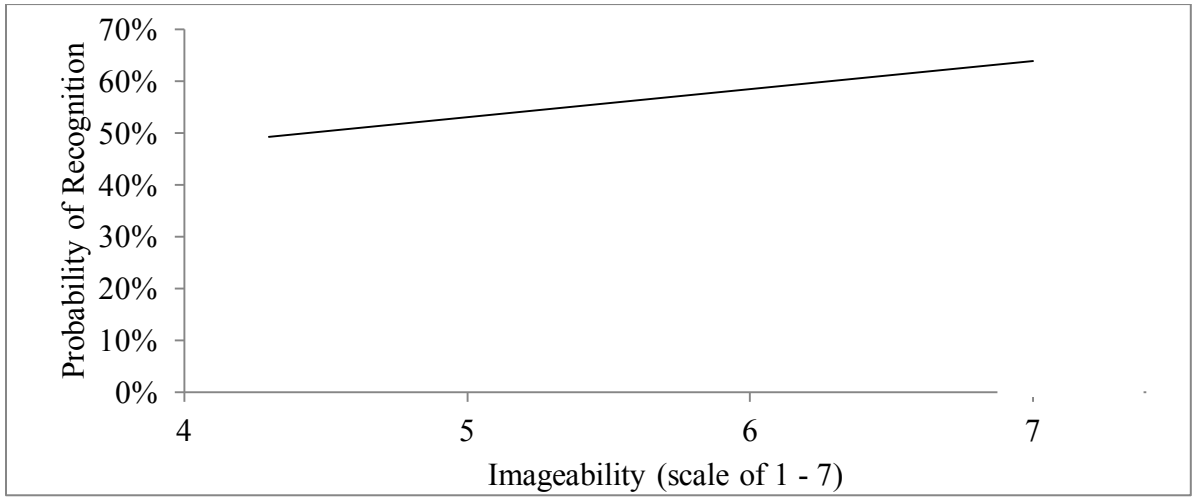

Figure 3.4. Model-based probability of recognition as a function of imageability (values ranged from 4.42 – 7). This illustration's abscissa is shortened accordingly.

Method of learning cannot be interpreted alone but requires specification of levels of word class. When considering nouns, the effect of inferential learning was to decrease odds of recognition by a factor of .45 relative to ostensive learning, which is a considerable reduction. However in the case of verbs, the negative effect of inferential learning was somewhat milder—verbs learned inferentially had only .63 times lower odds of recognition. This model shows ostensive learning was overall the better of the two ways to learn words. Similarly the effect of word class cannot be considered on its own without considering how these words were learned: when learned ostensively, verbs were .79 times less likely to be recognized than nouns. However when learned inferentially, verbs were 1.11 times *more* likely to be recognized than nouns. This interaction, illustrated in Figure 3.5, suggests that the noun bias among adults only holds for ostensive learning, and that when learned inferentially, a verb-bias takes effect.

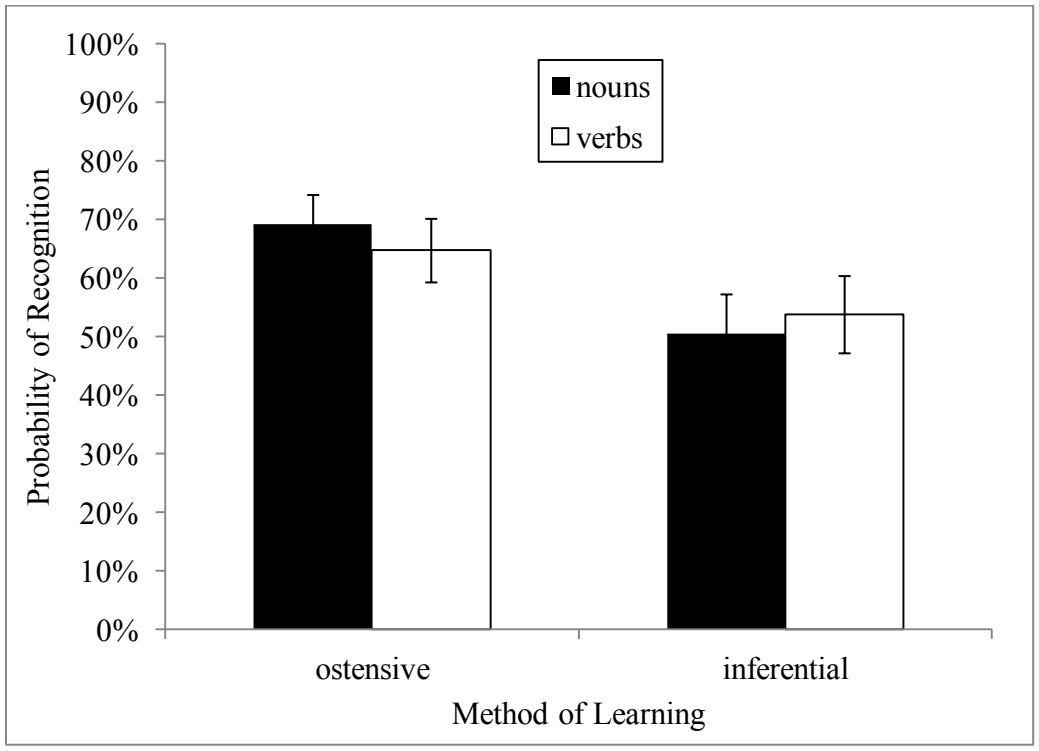

Figure 3.5. The model-specified effects of method of learning, word class, and their interaction on probability of recognition at five minutes. Error bars represent 95% confidence intervals.

An ongoing theme in this dissertation is on the role of word class in assessing word learnability. I wondered if some of the predictors in the five minute model were themselves correlated with word class such that their inclusion in this model would either increase or reduce the value of word class in the model. First I tested word class by itself: p=.04, odds ratio = .88. Then I tested word class with each of the other modeled factors, one pair at a time (always with word class) to see if any factors greatly increased or reduced the significance and effect size of word class when modeled together. With number of alternative interpretations, word class was brought down to p=.21 from p=.05, and its odds ratio was brought up to .92 from .88 (meaning the gap between nouns and verbs grew smaller). This means that number of alternative interpretations was correlated with recognition in a way that partially accounted for the word class effect. If word class were modeled without number of alternative interpretations included, the word class effect would seem to be due to word class entirely when number of alternative interpretations could potentially account for some of that effect (if it were a reliable source).

Only one factor improved the value of word class in this model—the interaction effect between method of learning and word class. When included, word class was brought down to p=.003, from p=.05, and its odds ratio was brought down to .79 from .88 (meaning with its inclusion, a larger disparity between word classes was found). This means that the word class differences would have been under-exaggerated if I failed to consider this interaction effect; accounting for the method by which nouns and verbs were learned caused word class as an effect to become more apparent.

One-week model.

Fewer participants were tested at one week than at five minutes; this section models learning with only the 56 individuals who were tested at one week delay. (I did not manipulate

delay when I tested 28 of the participants out of concern for data collection time limits.) With one-third fewer participants and approximately half the number of observations per participant, error variance is expected to be greater in these analyses, and power to detect effects lessened, relative to the general model developed earlier. Fewer factors should be modeled with smaller sample sizes.

The same eight factors initially entered in the five-minute model were also initially entered in the one week model, but the model was not significant. No single factor, modeled alone, was significant either. Modeling word class with name agreement in context images (the first-most significant, with the next-most significant factor) resulted in a significant model, Wald $\chi^2(2)=7.08$, $p=.03$. Modeling more than two or less than two resulted in non-significant models. Table 3.6 shows the two-factor model in which the effect of word class was significant, with verb recognition exceeding that of nouns, and the effect of context image name agreement was nearly significant, its trend showing greater name agreement was associated with greater recognition likelihood.[10]

I wanted to be sure the effect of word class did not owe to differences in depiction quality, but in testing this question I did not wish to enter more variables than the model could handle. I compared the effect of word class in its own model to its effect when modeled with each factor relating to depiction quality to confirm word class was an independent effect. The effect of word class, in its own one-factor model was nearly significant, $p=.06$, odds ratio = 1.16. I tested word class in a model with number of alternative interpretations to ensure its effect did

[10] The sample size should have been slightly larger for running this two-factor model. From the rule of thumb $N=50+8k$, I should have tested 64 rather than 56 participants had I known that I would test a two-factor model. Besides this, the number of models tested prior to establishing this particular model elevated alpha such that this model, when controlling for number of tests done, was no longer significant. I present it in spite of this to present a preliminary finding, a model to spark interest in future replication endeavors.

not owe to coincidental correlation with this factor. This model was significant, Wald $\chi^2(2)=6.79$, p=.03, with word class significant (p=.02, up from .06 when modeled alone) and having a similar odds ratio (1.23 from 1.16 when modeled alone). Thus name agreement and number of alternatives both improved estimated reliability and increased estimated effect size, though neither were themselves significant factors when modeled with word class. Similarly I entered imageability, number of occurrences, and method of learning, each in turn in two-factor models with word class to see whether the effect of word class was completely independent of these other factors. However these models were not significant. The story that emerges from these analyses is that word class was a real effect, context image name agreement likely was a likely effect, and by accounting for name agreement, the effect of word class could be measured and estimated more precisely (i.e., if one did not control for name agreement differences between nouns and verbs, the word class effect might not have been detected). The estimated odds ratio of 1.24 indicates verbs had an advantage over nouns when measured at one week. This surprising reversal from a noun advantage at five minutes to a verb advantage at one week has not to my knowledge ever been found. The factors of the "one week model" given in Table 3.6 are illustrated as probabilities in Figure 3.6.

Table 3.6

One Week Model of Word Recognition (Wald $\chi^2(2)=7.08$, p=.03)

			Measured values	
Predictor	p	odds ratio	min	max
Word class	0.02	1.24	noun	verb
Name agree-context	0.06	1.58	.08	1.00

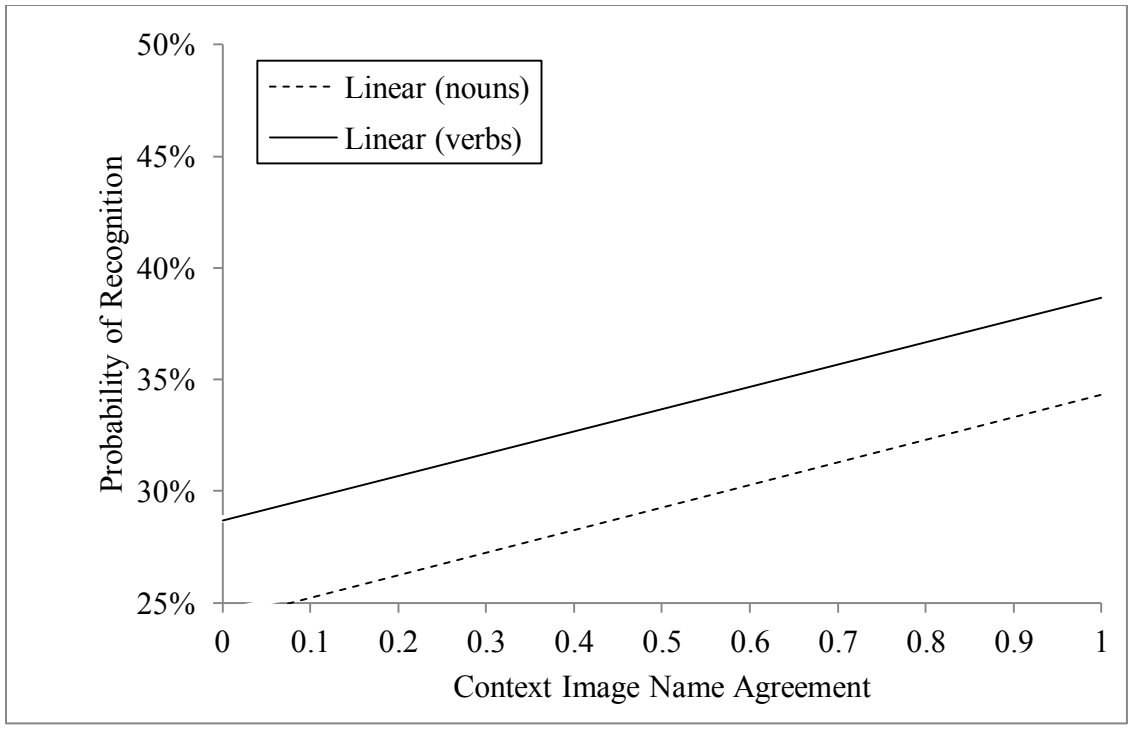

Figure 3.6. Model-based probabilities of noun and verb target recognition at one week as a function of context image name agreement values. The ordinate axis is shortened from 25 – 50% to provide a magnified view of these effects, and because chance performance was at 25%.

Significant at five minutes, it is curious that number of occurrences, method of learning, and imageability at one week were non-significant. That number of alternatives and name agreement bordered significance affirms the stability of their value to word learning.

General Discussion

Exploratory research ought to be accompanied by confirmatory research so as to ensure that findings based on exploratory approaches can be generalized and are not the product of spurious results based only on one particular sample of data (Simmons, Nelson, & Simonsohn, in press). This was in large part an exploratory study; confirmatory studies are still needed to confirm these effects, very especially at one week where reliability was weaker and the number

of participants was fewer. Some specific predictions were formed and tested in this study. Below I address these.

Answers to this chapter's questions

Question 1. Is there a noun bias among adults learning foreign vocabulary?

The answer to this question is yes at five minutes, and no at one week! The over-sized model indicated there was an interaction between delay and word class; this was corroborated with different effect directions in the five minute and one week models. The five minute model indicated noun superiority, while the one week model indicated noun inferiority. This is perhaps the most unique and memorable finding of this study, but it is elusive of an explanation. Implications and applications are presently completely lacking. Future researchers are called upon to suggest explanations and replicate this finding.

Question 2. Does the method of learning matter? Does it affect the rate of forgetting?

Uncertain to both questions. Method of learning significantly interacted with delay in the over-sized model, was significant at five minutes, but was not significant at one week. This pattern of findings suggests the method of learning does matter initially (at five minutes ostensive learning was superior to inferential recognition), but that its effect lessens with time. I predicted that inferential learning should result in less forgetting (better recognition performance) at one week than ostensive learning, but this effect could not be detected. This does not deny the possibility that my hypothesis was correct—word learning is a slow and incremental process and is almost never the result of a single exposure. Perhaps the un-measurable effect of method of learning may become multiplied (and measureable) over repeated learning opportunities. If targets were inferred (versus ostensively labeled) repeatedly from a

number of contexts, perhaps we would see that inferential learning is more robust to delays than ostensive learning. Further research is needed to test this.

Question 3. Is one method of learning better for verbs than nouns? Is one method better for nouns than verbs? Which hypothesis (parts before relations or reduction in referential ambiguity) is better supported by the results?

This study was designed to pit one hypothesis of word-learning—that verbs learned after nouns are learned better because prior noun argument knowledge enables verb learning—against another, more encompassing hypothesis—that removing ambiguity from the learning situation by any means enables verb learning, thus accounting for any benefit of prior noun knowledge in terms of reducing verb ambiguity. Initially I attempted to control ambiguity by adding marks and lines in verb images to symbolize movement and to make verb interpretations more likely. Even so, these doctored verb isolate images were not as well named as noun isolate images, nor were they rated as depicting their referents as well, and more alternative interpretations were offered of them than of noun images. Next I tested the role of ambiguity in learning with the predictors name agreement, goodness of interpretation, number of alternative interpretations. Although the number of alternative interpretations seemed the most promising measure of ambiguity and was significant in the over-sized model of word learning, it did not quite reach my significance threshold in either the five minute or one week models. The relative importance of referential ambiguity to this experimental paradigm appears weak. However, future investigators should consider the use of this measure as a way of quantifying referential ambiguity in either controlling or studying its effects.

By including number of alternative interpretations, ambiguity was more or less controlled. Under this circumstance I predicted, based on the referential ambiguity hypothesis,

there would be no interaction effect between word class and method of learning. Instead I found that the interaction effect was significant: among inferential conditions, verb learning was superior, but among ostensive conditions, noun learning was superior. This rather small interaction effect supports the parts before relations hypothesis better than the referential ambiguity hypothesis. Parts before relations is the hypothesis that (for underspecified reasons) knowledge of parts is critical to learning relation words, but the reverse is not so to the same extent. This is what was found—prior noun knowledge enabled greater verb learning than prior verb knowledge did noun learning. This could not have been due to differences in word learnability; the same words, when learned ostensively, revealed a noun bias under ostensive learning conditions.

However another way to interpret this interaction effect occurred to me: perhaps inferential learning is better suited to verbs than nouns. Verbs are relational concepts (verbs relate objects and a change of state to one another), and inferential learning highlights relationships between image pairs (the repeated and novel element in the context image is seen only in relation to the isolate image presented just prior). Perhaps inferential conditions prime relational thinking, thereby highlighting verbs more than nouns.

At one week there was no effect of method of learning, and no interaction between method of learning and word class. This suggests whatever advantages or disadvantages there may be for learning nouns before verbs, or learning inferentially rather than ostensively, their effects were highly transient. Elevating performance at the later delay might improve sensitivity to these effects. This might be done by testing after fewer days, by teaching fewer targets, or by teaching with more, unique isolate and context images rather than the same ones repeated many times. The trend toward significance of context image name agreement at one week signifies that

referential ambiguity may be an important predictor of learning success, but oddly its effect was not detected at five minutes.

Question 4. What are some other predictors of word learning?

The answer to this question depends on which of the presented models one subscribes to. The effects found in the over-sized model should be taken with skepticism as they may not generalize beyond this sample. The effects indicated in the one week model suffered the risk of Type I error in their interpretation. Therefore I first discuss effects found in the five minute model, and add a disclaimer before mentioning the others.

Number of occurrences.

Manipulating ostensive learning at both levels of this variable was a good decision; doing so allowed separating the effect of method of learning from number of occurrences. Very strangely, the greater number of times a target occurred within trials (2 times) was associated with a learning decrement at five-minute models. Perhaps participants found the redundancy in labeling within trials to be confusing. Given their task was to learn word meanings, the presentation of context images right after isolate images represents a form of massed learning. Massing should have led to better target memory relative to single occurrence conditions (i.e., not massing). That this was not the case was an intriguing observation that warranted follow-up investigation into learning words from more than one occurrence per trial. This is in fact the topic of the study presented in the next chapter.

Word Imageability.

Greater word imageability was associated with greater word recognition success at five minutes. This effect was even evident with number of alternative interpretations of images statistically controlled. Two implications come to mind. First, word learning researchers must

take note of target word imageability so that it does not become a confounding variable. Second, words that are known to be poorly imageable are less readily recognized and thus suffer from lower learnability; therefore more attention and training may be needed to learn such words. Knowing the difficulty of a word may inform instructors of the amount of time, attention and materials that are needed to teaching those words.

This concludes the list of other factors that affected word learning, as evident in the five minute model. There were some additional effects detected in the over-sized model and one week model, which I discuss next. Note, however, that these effects, discussed below, may not generalize beyond this data sample.

Name agreement.

This factor was not significant, but I mention it anyway because of the amount of effort invested in measuring it. It seems probable that illustration success is important for learning words from images. Yet name agreement was not a significant predictor of recognition in any model. Name agreement might have been a stronger predictor if images varied more in name agreement. Goodness of depiction and number of alternative interpretations seemed to be better measures of depiction quality in that they were more predictive of outcomes, and these were much easier to measure. Therefore name agreement measurement as a predictor of learning is not to be recommended for word learning research. However, future investigations may do well to manipulate referential ambiguity, or at least use materials with greater referential ambiguity variance, to find better evidence of its role. One way this could be done is by manipulating the number of distractor artifacts in each target image during learning, between participants.

Additional main effects, based on over-sized model.

Delay. This effect would most certainly generalize beyond this data sample! The odds of recognizing a word at five minutes were about 3.5 times larger than they were at one week, all other factors being even.

Word order. This effect suggests that among words learned ostensively, the verb-noun order led to superior recognition.

Alternative interpretations. As might be expected, as more alternative interpretations were given of an image, words were learned less well.

Interactions with delay, based on over-sized model.

English-as-first-language status. This finding, based on an uneven sample of 67 learners whose primary language was English, and 17 learners whose primary language was not English, showed a sharper decline in recognition over time for those whose primary language was not English, compared to English-first participants.

The experimental language. The effect of experimental language, not overall significant, suggests delay had a much smaller effect of reducing recognition when words were learned with Language B than with Language A. These two artificial languages were made up of the same stock of words but with re-assigned meanings. Therefore any differences in language learning must be attributed to specific item effects of certain words paired with certain meanings, rather than overall language differences in word sounds or spellings. This largest of interaction effects suggests two things: first, that it is important to account for language difficulty, and second, that certain words or languages may have different forgetting functions relative to others.

Word class. At five minutes, nouns stood higher odds of recognition, but at one week, verbs did. This amazing flip-flop of the word class effect over time eludes explanation. Do verbs have a slower forgetting function? Replication of this unexpected finding is needed.

Number of lists learned prior. This effect shows a slower forgetting function for the first list of words learned. Words learned in the second list were more forgotten at one week than at five minutes.

Caveats

Due to counterbalancing mistakes, one-third of the 96 words were only learned inferentially. The other two-thirds were properly counterbalanced between methods of learning between participants.

Order of words in ostensive trials was not counterbalanced by target; each target word was presented in the context of either the noun-verb order or the verb-noun order. Therefore the effect of word order could have been confounded by chance assignment of more learnable words to one of the two word orders.

CHAPTER 4: LEARNING NOUNS AND VERBS ACROSS SITUATIONS (STUDY 3)

Nitsch (1977) experimented on the effects of number of contexts in learning. She taught students novel meanings of novel words either by a) definition alone, b) definition with examples within one context (a cowboy story), or c) definition with examples from several story contexts (among them, the cowboy story). The effect of varying contexts on retrieval of word meanings in a unique story context was clear—students who learned across several contexts or situations performed best. This type of learning has been termed cross-situational learning, and has been a popular phenomenal topic of research in recent years, perhaps catching on from the work of Siskind (1996) and Akhtar and Montague (1999). From a cognitive perspective, cross-situational learning can be explained in terms of cue-overload theory (Watkin & Watkin, 1975) and stimulus sampling theory (Estes, 1955). Cue-overload theory says that remembering is more likely when multiple learning cues are provided because the additional cues tip the ratio of cues to targets in the memory-favored direction. Stimulus sampling theory says memory is a function of the number of stimuli that are present at the time of recall; by sampling from a variety of environments during learning, the likelihood of their presence during recall is higher and recall more likely.

Cross-situational learning is operationalized in the present study as learning the label for a target element repeated across a pair of context images in which non-target elements are not repeated, from short (2-word) phrases uttered with each context image to label elements in images. Cross-situational learning should reduce ambiguity for referents and thereby improve recognition performance. This should have been seen in the previous study in ostensive conditions with two occurrences per trial because those conditions were essentially cross-situational learning, but with one distinction: the first of each pair of "situations" (an isolate

image) perhaps conveyed its meaning too well so that the second occurrence may have been purely redundant and may have bored the learner. That study found a detrimental effect of cross-situational learning, which goes against all above-mentioned theories and findings of cross-situational learning benefits. In the present study I predicted a benefit of cross situational learning—learning a target from its use in two contexts—relative to a control method, herein dubbed one-situation learning—or learning a word's meaning from its use in a single context.

In the present study, "contexts" were scenes or images of concrete objects enacting concrete actions (many were the same context images used in the previous studies). By juxtaposing pairs of contexts, each sharing a single, common element (either an object or action element), cross-situational learning conditions were created, and learners were able to infer that the repeated component in each pair of images mapped onto the word element that was also common to both two-word phrase contexts. As a way of solving the ambiguity problem that ordinarily occurs in real-world, multi-word, multi-meaning contexts, an advantage for cross-situational over one-situation learning could be explained by the referential ambiguity hypothesis, the hypothesis that any reduction in ambiguity for word-to-meaning mapping will cause a corresponding increase in the probability of meaning recognition when given the word.

Cross-situational learning studies have become more common in recent years. However, very few cross-situational learning studies have aimed at uncovering a possible noun bias among child or adult learners. One notable exception is Piccin and Waxman (2007). They utilized the Human Simulation Paradigm, a paradigm which has recently become popular, especially for researching how learners use linguistic cues to learn words. Piccin and Waxman found adults and children were more successful at guessing nouns than verbs from beeps which replaced actual words in video dialogues. One shortcoming of that research is that in spite of measuring

and documenting a word class difference in imageability, this factor was not controlled or assessed and could possibly have confounded their interpretation of the overwhelming noun advantage they found. In the present study, the effect of imageability was considered.

The present study, a word learning experiment just like the prior study, differed from that study in several notable ways. First, the experimental language changed. The present study utilized Hebrew as the target language. From the previous study with nonsense words I found that utterance length was non-predictive of target recognition. I informally surveyed participants in the prior studies as to whether they thought the experimental language as a real or contrived language; about half thought it was contrived. Therefore I wished to extend my word learning research to a more realistic medium, a real language. I selected Hebrew because very few people are fluent in this language, worldwide, and because I found a Hebrew speaker for hire by which I obtained all auditory stimuli for the present study. I also wanted to employ a different voice than my own for experimental stimuli to reduce my role in the experiment from that of experimenter and teacher to that of experimenter only.

Other differences are notable as well. In the present study no words were uttered in isolation and no targets were presented as isolate images. Instead each trial was made up of the presentation of two context images, each accompanied by two-word phrases. Additional differences were that trials were presented only twice (in the prior experiments, each trial was repeated six times); the number of trials per block was reduced to only four (as opposed to eight trials per block in the prior study); trial orders were fixed (as opposed to trials presented in randomized order in each repeated block in the prior study); number of unique trials seen by each participant in the experiment was only 24 (as opposed to the number of unique trials being 48 before); the delay between learning and test was cut down to a mere 30 seconds and was not

manipulated (as opposed to a five-minute or one week delay); the test no longer relied on the English language and required only a mouse click on an image choice (as opposed to recognition being based on selection of the correct English word representing the meaning of targets learned from images using the keyboard); and only certain learned words were tested, one per trial, so that there were only 24 items tested (in the prior experiments, all 96 presented words were tested).

The two primary variables of interest in this study were "number of situations" (whose levels of learning were cross-situational and one-situation) and word class. I predicted one-situation word learning would lead to much less success than cross-situational learning for three reasons. First, in one-situation conditions, participants had little way of knowing that the target words were always spoken in common Hebrew syntax which is a noun-verb word order. In cross-situational conditions, Hebrew's syntax probably became somewhat apparent after a few trials because the shared element in both context utterances were always either uttered first or second, and were always either a noun or a verb, so participants could have quickly learned that nouns were always uttered first and verbs were always uttered second. Second, I predicted a disadvantage for one-situational learning because in these conditions intonation was the only clue for knowing how to parse words of each pair; some words may have been trickier to parse from phrases than others. For example in the phrase "tinok shotay" participants might have accidentally parsed "tinoksho tay." Such parsing errors could presumably lead to greater difficulty in recognizing target meanings given their words at test. Parsing was likely much more certain in the cross-situational condition because each trial was composed of a pair of phrases with one common, repeated word among each phrase pair. Identification of the common element would allow parsing it from its word pair, simultaneously defining the phonetic boundary of both

words. For example the phrase “atalef shotay” was followed with the phrase “tinok shotay,” allowing much greater certainty in correctly parsing “shotay” from each phrase. And third, research has shown that when a test situation differs from a learning situation(s), varied learning contexts are advantageous (e.g., Nitsch, 1977). In this experiment, the target images presented at test were different than the target images seen during learning. Therefore learning from a single context (e.g., one-situation learning) should be disadvantageous to word recognition from novel contexts.

I predicted an effect of word class based on the results of the prior study in which a noun bias was observed at the shorter of the two delays (five minutes). As the present study’s delay was fixed at 30 seconds, I expected to see nouns better recognized than verbs.

Based on the referential ambiguity hypothesis I predicted an interaction between number of situations and any of the measures of image referent ambiguity (name agreement, goodness of depiction, or number of alternative interpretations). I predicted measures of ambiguity would be more predictive of one-situation learning performance than cross-situational learning performance. In cross-situational learning, object- and action-identification of image elements should not fully depend on the referent-clarity of one of the pair of images—even if one image was ambiguous, the clarity of the other could make up for that ambiguity, allowing high accuracy of referent identification). Therefore referential ambiguity should all but disappear from cross-situational learning conditions, leaving image quality factors non-predictive.

I predicted a main effect of word class based on the observed noun advantage at the shorter testing delay in the previous study. Similarly, I predicted an interaction between number of situations and word class. Number of situations in the present study may be analogous to the method of learning variable in the previous study—both involved mapping words to meanings

with an inferential step. One-situation learning in the present study was only slightly similar to the ostensive method of learning in the previous study—both involved a bit of redundancy—but in one-situation learning, mapping word to meaning may have been far less certain for reasons mentioned already. At any rate the similarities between these study variables, together with the observed significant interaction between method of learning and word class found in the previous study, suggested to me there may be an interaction between number of situations and word class in the present study. I predicted noun learning would exceed verb learning in one-situation contexts, but that verb learning would be greater than noun learning in cross-situational contexts because cross-situation conditions may prime relational reasoning, highlighting relational referents.

Another endeavor in the present study was consider other measured factors as possible predictors of word learning, and to develop a model of word learning with those factors considered. Below is a recap of the main questions addressed in the present study.

1. Is cross-situational learning more efficient than one-situation learning?
2. Is there an effect of word class?
3. Does the word class effect vary by number of situations?
4. Do the measures of image quality predict learning differently by number of situations?
5. What are some other predictors of word learning within the present paradigm?

Method

Participants

Fifty undergraduate participants were recruited from a subject pool of students taking psychology or linguistics courses at the University of California, Los Angeles. To be eligible for this study, participants were required to not know any Hebrew. The first two participants were

run only to pilot the procedures; some procedures were altered in response to this dry run, and the data from those two participants were excluded from analysis.

The experiment proper sample included 48 participants. There were 37 females, and there were 38 native English speakers. The average age for participants was 21.5 years (range: 18-40, SD=5.0). The average of all participants' self rated proficiencies in English on a rating scale of 1 – 10 was 9.57 (SD=.96). Participants reported other known languages, ranging from none to five, and the average proficiency sum of each additionally known language, reported on the same 1 – 10 scale, across each participants' known languages was 18.49 (SD = 4.93), roughly meaning participants were fully proficient in one language and had acquired subsequent language(s), which, when added together was equal to 85% proficiency in a single additional language.

Design

The study was designed as two-way, mixed factorial. The first independent variable was word class manipulated within subjects at two levels—nouns and verbs. The second independent variable was the number of learning contexts (unique images) presented, manipulated at two levels between subjects: cross-situational versus one-situation (repetition) learning. One dependent variable, recognition, was measured on a binary scale (0=wrong, 1=correct). Repeated observations were nested within participants. Fourteen additional predictor variables, measured in the previous and present studies, were considered but not manipulated.

Materials

Targets.

Twenty-four context images were selected from the pool of 48 context images used in the previous studies based on their high target name agreement as measured in Study 1, and by avoiding verbs that could be considered transitive. These 24 targets are listed in Appendix A.

Measurements of these stimuli, obtained in Study 1 and provided in Table 4.1 below, were compared between nouns and verbs. Although nouns and verbs differed in phonemic and utterance lengths, utterance length was shown in the previous study not to affect recognition success. Besides these word factors, nouns and verbs also differed on only one other factor—target name agreement for images presented second in trials.

Table 4.1

Measures and Pair Differences of Nouns and Verbs in Study 3

		Nouns		Verbs			
	Factor	Mean	SD	Mean	SD	t	p
Words	utterance length	0.64	0.14	0.75	0.22	-2.29	0.03
	phonemic length	5	1.38	6.38	2.63	-2.26	0.03
	syllabic length	2.33	0.76	2.75	0.99	-1.64	0.11
Targets	familiarity	6.56	0.35	6.56	0.45	0.05	0.96
	frequency	3.56	0.89	3.89	0.99	-1.19	0.24
	imageability	6.78	0.25	6.74	0.25	0.43	0.67
Images	Goodness of depiction[a]	4.89	0.15	4.85	0.17	0.92	.36
	Alternative interpret[a].	4.29	2.07	4.87	2.85	-0.79	.43
	1st image	0.93	0.08	0.86	0.18	1.80	0.08
	2nd image	0.91	0.13	0.76	0.25	2.57	0.01
	isolate images	0.91	0.09	0.89	0.13	0.63	0.53

[a]These were measurements of isolate images, which were viewed only at test.

Auditory stimuli.

Language and speaker.

The speaker was a 38 years old native Hebrew-speaking male who was raised in Israel and moved to the U.S. as an adult. He spoke all words individually and in phrases at a normal speech rate as requested. He spoke all nouns in singular form; verbs were spoken in either masculine or feminine form according to the gender of the illustrated actor performing the verbs

in associated image. More verbs were spoken in the masculine than feminine forms. Phrases were spoken with natural articulation and sentential intonation.

Some Hebrew words were not suitable for this study because they were cognates of English words (e.g., *pinguin* in Hebrew means penguin). For these I substituted other Hebrew words that were recorded for this purpose. For example, I let the Hebrew word for angel, *malachit*, stand in for *pinguin*. Several substitutions were made, each substitution being from another Hebrew word of the same word class—nouns were substituted for nouns, verbs for verbs. This was done to maintain the target language's status as a completely foreign language to the participants.

Syntax. In Hebrew words are normally ordered subject, verb, object, as in English (personal communication with Hebrew tutor, March 2011), though more syntactic flexibility is allowed in Hebrew than in English (Jacobs, 2003). The speaker informed me that he produced all phrases in a noun-verb order.

Physical attributes. Recording was performed with a computer, microphone, and Audacity 1.3 (Beta) (sound recording software). Sound clips were edited to include an approximated 100-millisecond lag before speech onset and a 200-millisecond lag after speech offset to ensure that the complete word was uttered and that no soft or subtle word-parts were accidentally cropped during editing. Utterance lengths were measured to the nearest hundredth of a second. Table 4.1 (above) provides the averages for noun and verb utterance lengths.

Numbers of phonemes and syllables were counted for all auditory stimuli. Independent samples t tests showed significant differences in the lengths of spoken Hebrew nouns and verbs. Although number of syllables did not significantly differ, $p=.11$, verbs were generally longer in phonemic length, $t(46)=-2.26$, $SE=.61$, $p=.03$, and had longer utterance lengths, $t(46)=-2.29$,

SE=.05, p=.03, (all tests two-tailed). In Study 2 I found utterance length was unrelated to word recognition, so these word length differences should not confound interpretation of any word class effect.

Images.

Sources. A third (24) of the 72 context images used in the learning phase of this experiment were also used in the two previous studies. These and the remaining two-thirds of images were mostly collected from the Internet but some were drawn by hand by either of two artistic research assistants. All images in the learning portion of the present study were context images—images meant to convey two concepts, always an actor performing an action. Finally, six additional images were found and used for training purposes only.

Physical measurements. The heights and widths of image stimuli were measured with a mouse using a pixel ruler (freeware); these measurements were used to check that there were no systematic size differences among image stimuli. Heights and widths were summed to yield a composite measure of each image size. The composite sizes were grouped into three groups based on their teaching function—images positioned second in trials, images positioned first sharing a noun with an image positioned second, and images positioned first sharing a verb with an image positioned second (with 24 images measures in each group). Images positioned first in trials to teach nouns (M=933 pixels, SD=147) were no larger than those used to teach verbs (M=991 pixels, SD=161), independent t(46)=1.34, p=.19, demonstrating no learning advantage for nouns or verbs due to image size.

The learning program.

The same laptop computer and stimulus presentation software used in the previous studies was used again.

Learning.

The same terms used to describe programming of the presentation of stimuli in the previous experiments are used again here. Each learning trial was composed of a pair of events, the context images. Within each block, four trials were shown once each. Each block was immediately repeated once, presenting those trials in the exact same order. The block and its repetition made up a segment, which lasted about 48 seconds. Participants saw a total of 6 segments.

Trials. Each trial was composed of two context image events presented serially for three seconds each. At each image onset a two-word phrase sound clip was played. Figure 4.1 demonstrates with flow diagrams two examples of cross-situational learning trials. Participants assigned to cross-situational learning would have seen one or the other of the flow diagrams illustrated in the figure, below. Each pair of context images depicted a total of three referenced elements, and sound clips referenced these three: a noun and two verbs, or a verb and two nouns. Only one target (a noun or a verb) was observed twice among the two images, allowing cross situational learning of that word's meaning. With images symbolized as letter bigrams, this learning series was AC (first image), AB (second image) for learning a noun across contexts, or DB (first image), AB (second image) for learning a verb across contexts. Underlined letters indicate the common referent between image pairs.

Figure 4.1. Pairs of context images. Each cross-situational learning trial was composed of a pair of context images linked by a common element (either a noun or a verb). The upper left panel should show an image of a king who is frowning (I hand drew this image rather than presenting the actual experimental image used; that image was copyrighted) and the upper right panel shows an image of another king who is typing, teaching the word for king. The bottom panels show an image of an alien who is typing and a king who is typing, teaching the word for typing. The remaining artistic contributions and adaptations above were freely given by Goldie Salimkhan and Kay Lee.

Of those participants assigned to cross-situational learning, half saw each trial begin with an image sharing a noun or a verb with the second image; for the other half of participants assigned to cross-situational learning, each trial began with images sharing a verb or noun with the second image on each trial (the word class of the shared component per image pairs was different between these two groups of participants). Between these participant groups, images

viewed first in trials were different, but images viewed second remained the same. In letter symbols, all participants groups viewed image AB second, but half of those assigned to cross-situational learning viewed image AC first while the other half viewed image DB first. In this way, participants learned, across situations, either a noun or a verb from image AB.

All participants assigned to one-situation learning saw the same set of images, set "AB" using the above-given letter rubric, as those assigned to cross-situational learning, but none saw image sets AC or DB. Instead each image of set AB was seen twice in a row. Figure 4.2 demonstrates this with a flow diagram. A sound clip referenced both targets, and was played at the onset of the image, and repeated with the image's repetition. The learning series progressed as AB, AB.

Figure 4.2. A pair of (identical) context images. Each one-situation learning trial was composed of two of the same image. A few milliseconds of white screen intervened between these images to create the experience of seeing two images rather than one. Images contributed freely by Goldie Salimkhan.

Blocks. In each block four trials were presented and repeated. Four orders were created and used between participants, but only one was assigned and used for each participant for the presentation and repetition of blocks of trials. The four orders of these four trials were counterbalanced between participants to control sequence and order effects.

Testing.

A switch to using image choices rather than word choices was a major improvement from the previous study. Word learning may progress gradually in a way that cannot be measured well with a test that asks for translated products. A word's meaning may not lend itself well to translation until it reaches a vocabulary-mature point in its development. The gradually maturing hypothesis of a word's meanings may not yet be as nuanced as any specific translation into the native tongue. As a hypothetical example, with limited exposure and before fully understanding that *rofe* means king in Hebrew (the mature vocabulary state), a learner could hypothesize *rofe* means something like royalty, prince, kingliness, highest class, or anything evident in the "rofe" situation(s) where learning has so far occurred, but the learner might not have understood *rofe* to mean king exactly in her native language yet. However as the learner becomes further exposed to more examples, all false definitional aspects eventually should be pruned away from the learner's working definitional hypothesis leaving behind an accurate definition, one that may be translate-able in the learner's native language. Learners who are not yet at the translation-ready stage of definitional development might have trouble picking a target's translation from a line-up of choices, but should have far less trouble identifying an image example because images often depict less nuanced versions of element meanings than words.

The learning assessment was a multiple-choice image recognition test. The test was to select each target word's correct meaning from four image choices (chance performance was 25%). Two of the four image meanings co-occurred within one image of the set of context images AB during the learning phase—as did the other two test choices, following the test design logic given in the previous study. This was to ensure performance could not be based only on associating a target word to its entire context image (i.e., participants had to know which target

word mapped to which image element). There were always two nouns and two verbs among the options; the presence of two noun options helped to ensure participants' performance was not based only on learning a target's word class. Tests were given after each learning segment and 30-second distraction task. Each test contained four test items. Six tests followed the six segments. In total 24 words were tested per participant.

Targets.

For participants assigned to the cross-situational learning condition, only meanings learned across situations were correct options. For those assigned to the one-situation learning condition, only one meaning per trial was tested as the correct option for half of these participants, and the other meaning per trial was tested as the correct option for the other half of these participants. For all participants, half of the tested elements were nouns and half were verbs. In other words, half of all participants were tested on one of the pair of elements in each image of set AB, and the other half were tested on the other element of these pairs. Thus each participant was tested on recognition of 24 words, and across participants, 48 words were tested. Target location was randomly chosen and counterbalanced between the four locations on the screen across test items to prevent location bias. The locations of foils were also randomly distributed to disguise the relationship between foils and targets.

Foils. Among all test items, the target image of one test trial was presented again as a foil option in one other test trial. In this way every image option was presented twice during testing. No element pair (from image set AB) was presented with another element pair more than once as choices at test. For example, if the elements of the image of a baby drinking were positioned with those of penguin painting as choices on one test item, baby and drinking might be positioned with king and typing (but not penguin and painting) in a later test item, regardless of

which was the target. This was done to discourage participants from using a process of elimination strategy based on their performance on prior items.

Procedures

Participants were run by one of four research team members. Participants began the experiment by completing a consent form and biographical data sheet that asked their age, sex, known languages, and proficiencies in those languages. Next the experiment proper was run, followed by an image naming task. Finally, participants were debriefed and thanked, and 1 credit (for one hour of participation) was awarded. The experiment usually lasted about 30 minutes. Procedural details of the experiment proper and naming task follow.

Experiment proper

Training.

Instructions presented on screen informed participants that they would be presented with images and short Hebrew phrases describing those images, and that the stimuli would progress at a rate of one image every three seconds. Participants were presented with two learning trials (four images), presented back-to-back, similar to the experiment proper (but with two trials instead of four, and without repetition). Then instructions immediately appeared on screen to select the correct target meanings of words just learned (but in the experiment proper, learning and testing were separated by a 30-second interval). After completing two test trials (whose choices were all four image elements seen just earlier in this training) with the mouse, participants were asked to explain their task in the experiment. The experimenter corrected any response that was not similar to "learn the meanings of the words." Then the experiment proper began.

Learning.

Learning began immediately after training was completed. Participants were presented with six learning segments interleaved with filler tasks and tests over immediately preceding segments.

Filler task.

After each segment, a filler task was given to participants to work on for 30 seconds. This filler was selected to prevent auditory rehearsal of words just heard. Participants were instructed to read the first (or next) question on a sheet of paper and answer it in writing on the paper. If they finished writing their response before 30 seconds had passed, the experimenter instructed the participant to continue writing until 30 seconds had passed, at which point the participant was stopped from writing. The questions asked in the filler task were made to be interesting and thought-provoking to discourage rehearsing words just learned. Responses to these questions were not analyzed.

Testing.

Following each 30 second filler task, a recognition test was given. Each test was only four items long, testing recognition of one target per trial of the four trials per block.

Naming Task.

After the experiment was finished, participants were given context image identification task so that I could measure name agreement among context image elements not already measured in Study 1. To avoid having participants name context images seen during the experiment proper (because they might have gained more or less of a naming edge depending on level of number of situations to which they were assigned), each participant viewed and named the 24 images they had not seen during the experiment proper. In this task, participants were

asked to type two or more words to describe each image (each of which depicted one noun and one verb). I randomly selected 24 participants' responses, 12 who named one half of the images presented first in trials, and 12 others who named the other half of those images. Each participant's responses were coded by just one of three available coders who were instructed to make judgments of name accuracy using three possible codes: 0=wrong, 0.5=partially correct, 1=correct. The three coders coded six participants' data in common, which accounted for 25% of the coded data, to establish a measure of inter-rater reliability. Krippendorff's alpha (a measure of inter-coder reliability) was found for each of these six participants. These six alphas averaged .84, exceeding the .80 threshold criterion suggested by Krippendorff (2004) for drawing "safe" conclusions.

Results

Data were analyzed with logistic regression, a method appropriate for data with a binary outcome variable (correct versus incorrect choices). With logistic regression I could see which, if any, factors of targets, images, or participants were predictive of recognition, and could control those factors if necessary while analyzing the effects of other factors. However one disadvantage of using logistic regression in this instance was that model testing with a sample this size (N=48) must be limited to small models, two factors at most. In the process of exploring the data, I developed a small model of word learning within the given paradigm.

Surveying the predictive worth of measured factors

A number of factors were considered for logistic analysis, several of which were first measured in Study 1. Using the same exploratory tactics employed in Study 2, I initially tested these 18 factors in individual models. Results and descriptive statistics are displayed in Table 4.2, below. An alpha criterion of .05 was used for determination of significance.

Table 4.2

All 18 Factors Tested with Individual Models

Factor Types	Factors	Mean (SD) or % cases	Wald χ^2	p
Independent	Word class	50.0% nouns	6.57	.01
	Number of situations	50.0% within	22.89	.00
Image	Goodness of depiction	4.87 (.16)	2.47	.12
	Alternative interpretations	4.57 (2.45)	3.62	.06
	Name agree, 1st imag	0.90 (.14)	0.57	.45
	Name agree, 2nd imag "AB"	0.84 (.21)	5.45	.02
	Name agree, isolate	0.90 (.11)	0.01	.92
Hebrew Word	Utterance length	0.70 (.19)	0.4	.53
	Syllable	2.54 (.89)	1.53	.22
	Phoneme	5.69 (2.17)	0.05	.82
English Word	Familiarity	6.56 (.39)	1.59	.21
	Imageability	6.76 (.25)	0.51	.48
	Frequency	3.72 (.94)	3.11	.08
Participant	Age	21.53 (5.01)	0.69	.41
	Sex	77.1% female	2.13	.15
	English 1st language	78.7% Eng 1st	1.63	.20
	English proficiency	9.57 (.96)	0.51	.22
	Total proficiency	18.49 (4.93)	3.01	.08

Main effects.

Number of situations was significant, model Wald $\chi^2(1)=22.89$, $p<.01$, favoring cross-situational learning over one-situation learning. Word class was also significant, model Wald $\chi^2(1)=6.57$, $p=.01$, with a learning advantage for nouns over verbs. These were therefore re-introduced into more developed models.

Among sets of images measured on name agreement, only name agreement among image set AB was significant, model Wald $\chi^2(1)=5.45$, $p=.02$. For purposes of model development I would only have taken one of these measures of name agreement forward with model development anyway because of inter-correlations between name agreement indices, all $r > .40$.

Thus name agreement for context images presented second in trials was tested again in more developed models. Number of alternative interpretations bordered on significance, Wald $\chi^2(1)=3.62$, $p=.06$, so I also tested this factor in later models.

Interactions.

Interactions were tested in models including each pair of factors and their interaction factor. The predicted interaction between name agreement and number of situations was not observed, $p=.99$. Also not observed was an interaction between number of situations and goodness of image depiction, $p=.96$, nor between number of situations and number of alternative interpretations, $p=.64$. Also not observed was the predicted interaction between number of situations and word class was, $p=.93$. Thus contrary to prediction, cross-situational learning did not reduce the importance of high name agreement in learning from images, nor did it alter the word class effect in effect in one-situation learning conditions.

Model development

Having determined which individual factors might affect learning, I next entered predictors together in models to confirm that the above findings held true even when other significant factors were held constant. I entered name agreement of image set AB (but not name agreement of images presented first in trials because that was not predictive, and because the two indices were correlated), word class, and number of situations into a model because these were all reliable, individually. This model was significant, Wald $\chi^2(3)=29.55$, $p<.001$, with word class ($p=.003$) and number of situations ($p<.001$) significant, but name agreement was not significant in this model, $p=.15$ so I removed it. Also given the sample size, really no more than two factors should be modeled simultaneously.

Next I wanted to know whether the effect of word class might owe to differences in name agreement, number of alternative interpretations, or imageability. To address this, I tested word class with each of these factors, one by one, in two-factor models (to keep models small). First I modeled number of alternative interpretations (which bordered significance in its own model) with word class. Alternative interpretations was itself not a significant component, $p=.11$, but word class remained significant, $p=.02$, meaning even when number of alternative interpretations was controlled, word class continued to explain performance. Next I did the same thing with imageability to confirm that the effects of word class remained when imageability was controlled. Word class remained significant, $p=.01$, but imageability was not, $p=.38$, meaning imageability could not account for performance beyond word class, but word class could account for some performance variation even when imageability was controlled. Finally I tested word class and name agreement of image set AB. Neither factor, word class ($p=.07$) or name agreement ($p=.13$), was reliable, which was indicative of a correlation between them and with the outcome variable. Both factors formed significant models with number of situations, and word class was the slightly more reliable of the two models.

That name agreement could reduce the reliability of word class to insignificance indicates name agreement, too, may be an important effect, and that the effect of word class could potentially be attributed to differences in name agreement. I tested name agreement with imageability and number of alternative interpretations to see if name agreement, too, explained some outcome variance beyond imageability or number of alternative interpretations. Although it remained significant beyond imageability ($p<.05$), name agreement was not significant when controlling for number of alternative interpretations ($p>.05$). On this, and because word class was slightly more reliable, I present the model of word learning without name agreement as a factor

in Table 4.3. Note, however, that the model with name agreement was nearly as reliable. These effects on the probability of recognition are illustrated in Figure 4.3. One word of caution about this model is that even with only two factors modeled, this model violates the rule of thumb N=50+8k which prescribes that with the given sample size (N=48), modeling two factors would require increasing the sample size by another 18 participants. In this sense, the model itself is a bit "over-sized," and may not generalize beyond the present sample of data.

Table 4.3

Model of Word Recognition in Study 3 (Model Wald $\chi^2(2)=27.37$, $p<.001$)

			Measured Values	
Factor	p	odds ratio	min	Max
Number of situations	<.001	2.82	one	cross
Word class	.01	0.72	noun	verb

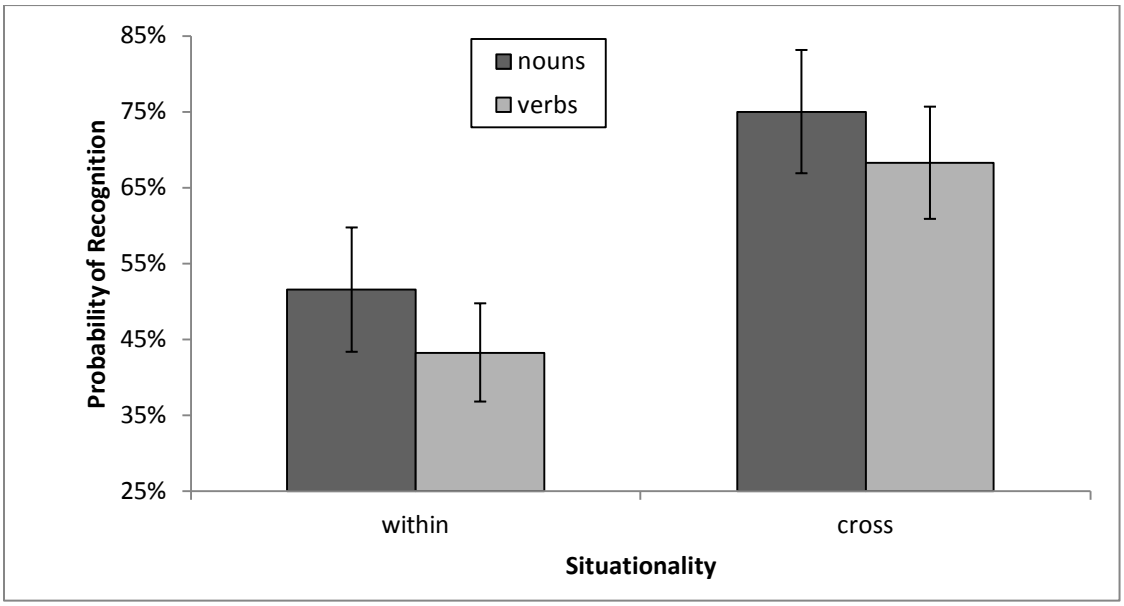

Figure 4.3. The model effects of number of situations and word class on probability of image recognition. The ordinate axis is shortened from 25% - 85% to magnify effects and to mark chance performance at 25%. Error bars represent 95% confidence intervals.

Cross-situational learning led to 2.82 times greater odds of recognition than one-situation learning which is a large effect. Verbs stood .72 times worse odds of recognition than nouns, a moderate effect size, which is very near to the odds ratio of word class found in the previous study (.82). Put differently, nouns in the present study had 1.39 times higher odds of recognition than verbs. The noun bias observed at five minutes in Study 2 was thus replicated and generalized to this study's paradigm and shorter delay, 30 seconds.

Discussion

Eighteen predictors were tested; only two were significant when controlling for one another—number of situations and word class. The questions posed at the beginning of this chapter are addressed below.

Question 1. Is cross-situational learning more efficient than one-situation learning?

The answer to this question is very certainly affirmative. The great advantage of cross-situational learning relative to one-situation learning could have been for any of the three reasons I cited for predicting this effect: by providing indicators of the syntax of the target language, by allowing more accurate parsing of words, or by providing more meaningful learning situations (or by any combination of these reasons). More nuanced experiments are needed to know why cross-situational learning was advantageous in this study.

The large effect of cross-situational learning over one-situation learning found here agrees with Nitsch (1977), in which it was found that a more varied learning context led to greater performance on word retrieval in a new context. By re-defining contexts as simple line drawing images housing target referents, I have shown that varied contexts promote better word learning even when those contexts are given very briefly (three seconds). Cross-situational learning is arguably the most common kind of learning in real-world settings, and by this study's

analysis, fortunately so. The cross-situational learning conditions in the present experiment likely improved learning by making targets more meaningful. By incorporating multiple, varied learning situations in language lessons, language instructors may provide an efficient way for learners to acquire new vocabulary. Review is important in any learning situation, but might be made more effective with the use of multiple unique contexts or examples.

Question 2. Is there an effect of word class?

Yes, nouns were better learned than verbs in this study. The word class effect found in the prior study was replicated in this paradigm and at this even shorter delay. This noun bias was not due to differences in noun and verb imageability (and by extension, its correlates), or number of alternative interpretations, as evidenced by maintenance of significance values of word class when these two factors were modeled (and thus controlled) with word class. It was less clear if the effect of word class was separate from the effect of name agreement, or accounted by it. If one accounted for the other, the more likely direction is that word class accounted for name agreement variance because, when modeled together, the former more closely approached significance than the latter. Whether the effect of word class found here generalizes to all concrete nouns and concrete verbs remains an open question.

Question 3. Does the word class effect vary by number of situations?

Nouns were learned better in cross-situational learning, but they were also learned better in one-situation conditions. The word class difference was constant at both levels of number of situations, therefore these nouns were more learnable than these verbs. My prediction that the word class effect would disappear in cross situational learning did not bear out. The referential ambiguity hypothesis, on which this prediction was based, posits that ambiguity reduction promotes learning. I assumed cross-situational learning would remove so much referential

ambiguity from learning situations that it would place nouns and verbs on par, both devoid of ambiguity under such conditions. That cross-situational learning in no way reduced the word class gap from that seen in the more ambiguity-filled one-situation conditions shows that cross-situational learning provided equal gains for nouns as for verbs and did not affect the learning gap seen in one-situation conditions.

Question 4. Do the measures of image quality predict learning differently by number of situations?

I predicted that measures of image quality would be more predictive of learning in one-situation conditions. No interactions were detected between number of situations and any of the candidate measures of image quality—name agreement among image set AB, goodness of depiction, or number of alternative interpretations of images. Thus no evidence was found to suggest differences in the predictive values of these variables across levels of number of situations.

Question 5. What are some other predictors of word learning within the present paradigm?

Besides number of situations and word class, no other factors were definitively established as predictors of word learning. Name agreement was a close competitor with word class; the two both produces significant models on their own, and when modeled together both reduced the other to insignificance. The decision to include word class rather than name agreement in the developed model was made not without some uncertainty, but because word class was the more significant of the two when they were modeled together. Perhaps if more data

were collected, name agreement among image set AB would be seen as predictive in addition to word class.[11]

[11] Why was name agreement among image set AB predictive, but not name agreement among images positioned first within trials? One possible explanation is that these latter values were less accurate name agreement measurements. Name agreement of image set AB, obtained from Study 1, were based on the average of 10 participants' responses; each response was coded by six coders, and the average of these six codes per response was taken as a measure of each response. Thus in Study 1, each word's context image name agreement value was the result of averaging 60 measurements (10 multiplied by 6). Name agreement for images positioned first in trials in the present study were calculated by averaging over only the participants (each word was measured as the average of response accuracy of 12 participants), each coded by a single coder. Thus each name agreement value for images presented first in trials was the result of averaging over only 12 measurements. This might have been sufficient if coders were "spot on" with each judgment but coders were required to use only three code values (0 / .5 / 1) to assess name agreement in spite of the construct really existing on a ratio scale. Perhaps name agreement of images presented first in trials, measured by averaging over fewer values, was a blunter instrumental measure, too blunt to be useful. An alternative explanation for name agreement among image set AB being more significant than images presented first in trials is that participants learned from, or learned to learn from, images of set AB only; they may have noticed that in each trial, image AB held the key to solving the word-to-referent mapping problem, and may have attended more to this image because of this. If this question is interesting, future researchers could test whether a first, clear example relieves the necessity for subsequent clear examples, or if instead learning is the sum product of example clarity of all examples seen.

CHAPTER 5: LEARNING NOUNS AND VERBS

The two goals of this chapter are to summarize important findings from the studies I ran and to give an account of them. The findings of this dissertation are summarized and framed in terms of the overarching goals of this dissertation: to see if and why nouns may more learnable than verbs among adult foreign word learners, and to know what additional factors (other than word class) may account for word learnability.

Whether there was a noun bias

The measurements conducted in Study 1 were suggestive of factors that may account for a possible noun bias among adults, and word learnability in general. These were differences between nouns and verbs in imageability, goodness of depiction, number of alternative interpretations, and name agreement. I suggested that factors which might account for word learning should contribute toward name agreement as a proxy for word learnability from such images. Imageability accounted for name agreement above and beyond familiarity and frequency which made it the most likely candidate of its correlates to account for word learnability. Likewise goodness of depiction and number of alternative interpretations contributed toward name agreement.

However I found that name agreement was actually a weak proxy for word learnability: although it showed a predictive trend in Study 2 (one week model) and Study 3 (in its one-factor model), it could not reliably predict word learning when word class was controlled. Word class was the only factor that was significant in all models developed, but the direction of its effects was not always the same (there was a verb advantage in the one week model developed in Study 2). One of the major goals of this dissertation was to discover if a noun bias exists among mature learners of foreign vocabulary. On these results a noun bias may very well exist but its presence

seems at least partly attributable to differences in name agreement, imageability, and / or other correlated factors.

Why a noun bias

Another major goal of this dissertation was to try to account for the effect of word class if possible. I found several factors that may partly account for word learnability: name agreement and number of alternative interpretations—both measures of depiction quality—were correlated with word class. When I regressed word class with these predictors on word recognition, number of alternative interpretations (Study 2) and name agreement (Study 3) slightly decreased the reliability of word class and its effect size. In other words, a small part of the reason verbs were harder to learn seemed to be their poorer depictions in images. This implies the noun bias among adults learning foreign words from images may due in part to depiction difficulties. Animation or video footage, or even better images might have reduced alternative interpretations and improved name agreement measures, perhaps with a corresponding improvement in learning and reduction in the effect of word class. Media that produce higher measures of name agreement may be useful to foreign language instruction. But media quality could not completely explain the observed noun bias. In the absence of a clear reason for the word class effect, I must turn to the noun bias that presides over early first language learners, and suggest that the noun bias lives on into adulthood.

Exceptions to the noun bias, and post hoc accounts

The noun bias did not appear across all conditions. Two conditions were associated with greater verb learning. In particular, verbs stood significantly higher odds of being recognized than nouns when tested at one week, and stood numerically higher odds when learned inferentially. That more verbs than nouns were recognized at one week was an unexpected

finding—perhaps the first of its kind. This novel finding suggests the forgetting rate for verbs is slower than that for nouns. This is an exciting result because it is a novel finding. However many analyses were run in an exploratory manner with this data, and the significance of the word class effect at one week could not survive the required alpha inflation as a result of multiple analyses performed. Therefore future research is needed to confirm this preliminary finding, controlling for word learnability.

One post hoc theory that can be offered for the noun advantage at short delays and verb advantage at the long delay is that verb learning is less-guaranteed but more immune to forgetting. One way to test this theory is to test noun and verb recognition, holding initial learning certainty constant (perhaps as indicated by ceiling performance at immediate test), then either testing all words again at several delays, or testing subsets of words at punctuated delays, to see if performance declines for verbs at the same rate as it does for nouns.

The other variable level that caused verbs to be better learned was inferential learning. In Study 2 I hypothesized this was due to the method's highlighting of relational aspects; however, when I tested this hypothesis in Study 3, it was not confirmed. In Study 3 nouns were better learned than verbs, and this difference did not vary by level of number of situations. The presumed relational highlighting of verbs in cross-situational learning turned out to be unlikely. In retrospect, I believe the relational highlighting hypothesis was too simplistic to explain the verb advantage in Study 2's inferential condition at five minutes.

Another explanation of this verb advantage can now be offered. Inferential learning, which occurred in the second image event of each trial, required mapping a novel word to a novel element. The saliency of the novel element would have been a function of the saliency of which element was repeated. That is, the novel element in the second image would only have

been as clear to learners as it was clear which element was repeated from the first image. On first appearance this sounds like a trivial task, but there may have been some real ambiguity about which element in image pairs was repeated. For example, to learn "penguin" by inference, participants first saw an image of a man painting a wall, followed by an image of a penguin painting a picture. The goals in these two actions are really quite different, and only sound the same upon verbally describing them. Only in a very broad sense of the word "painting" was painting actually repeated. This task's difficulty may have been multiplied when the novel element was not very apparent in the image.

Also, although all of the verb stimuli could, in English, be spoken intransitively and make sense (e.g., "A penguin paints."), their visual depiction often required the existence of a patient (e.g., painting a wall, painting a picture). In this painting example, there were actually at least two *novel* elements presented in the image of a penguin painting: a penguin, and *painting a picture*. Across studies the presence of a novel actor, and sometimes novel patient, in the second images could have made cross-situational identification of verbs less certain. In graphical depictions of referents, there may generally be greater variance from one image to the next among verbs than among nouns. Inherently goal-directed in meaning, verbs in this sample may have shared less in common between their image pairs.

Greenfield and Alvarez (1980) showed that when the number of unknown referents is reduced, recall was more likely. There would have been fewer "unknowns" in isolate images of nouns than of verbs, given verb depictions require actors and sometimes patients. Gleitman et al.'s (2006) statement that there is greater "surface variability in how verbs get realized . . . within and across languages" (p. 32) begins to take on greater meaning in light of the above considerations. Verbs with the same name and definition are frequently used to refer to actions

with unrelated goals, actors, patients, and instruments. I shall call this account the "surface variability hypothesis". This hypothesis can be used to explain why nouns, generally more learnable than verbs, were not as well learned under inferential conditions because noun inference depended upon establishing verb meanings beforehand, which, in this account, was not guaranteed.

Accounting for the noun bias

The surface variability hypothesis also presents a likely explanation of the noun bias among learners of novel, concrete vocabulary in general—namely, that verb situations naturally arise with great variance in terms of the goals, actors, patients, and tools involved, making identification of verbs from these situations less certain.

One of the goals of Study 2 was to differentiate between two other, seemingly competing explanations offered in Greenfield & Alvarez (1980) for a possible noun advantage among mature foreign language learners. They found learning parts before relations aided learning, and also that reducing unknowns (i.e., ambiguity) aided learning. It can be inferred from their study that learning parts before relations was successful because it reduced ambiguity. In other words these hypotheses may be viewed as complementary rather than competitive. In Study 2, learning parts before relations aided learning of relation words more than relations before parts aided learning part words, in support of the parts before relations hypothesis. In the real world, word learning situations are rarely packaged neatly into ostensive situations involving such obvious mapping of one referent with one word (although these may be common in the language classroom). In ambiguous circumstances the referential ambiguity hypothesis, offering a broader scope of explanation, may be more useful. Although I offered the surface variability hypothesis as an alternative interpretation of the verb advantage in inferential conditions, I do not see these

three hypotheses (parts before relations, reduction in referential ambiguity, and surface variability) as mutually exclusive of one another.

Cross-situational learning in Study 3 was not a very good analog of inferential learning in Study 2, in retrospect. In cross-situational learning, it was the repeated element across two images, and not the novel element in the second images, that was tested. Thus the noun bias seen in cross-situational conditions trials is not inconsistent with the explanations of the verb advantage for inferential learning offered above.

Framing this dissertation within its theoretical context

In the introductory chapter to this dissertation, a set of theories was introduced, each one aiming to account for the noun bias phenomenon observed in early children's vocabulary development. Would the findings of this dissertation support or controvert their transplantation onto the possible noun bias phenomenon observed among the adult samples in this dissertation? Throughout this dissertation I have supported the referential ambiguity hypothesis as a viable explanation for my results. Now I consider some of the other presented views.

The natural partitions – relational relativity hypothesis proposed by Gentner (1982) is arguably the best-cited explanation of a noun bias. In accord with this theory, noun context image name agreement was higher than that of verbs. The natural partitions aspect of Gentner's theory accounts for this—that nouns are easier to identify because they are more easily partitioned from their environmental context. The relational relativity aspect of Gentner's theory is that verbs are difficult to acquire because they label a varying fragment of a change or action scene—which aspect is referenced by the label is not given in the situation itself. Unfortunately the relational relativity aspect of Gentner's theory does not apply well to verb learning in this dissertation because all verb stimuli were developed based on English words, and all participants

knew English well. Therefore participants' assumptions regarding how to segment actions or how to set semantic boundaries on verb meanings would in most instances have been correct.

Naigles (1990) and others have suggested morphosyntactic complexity might account for the noun bias. Verbs tend to allow and require more morphological inflection than nouns across instances and languages. In this dissertation, no morphological inflections occurred, whatsoever. In Study 2 labels were counterbalanced between nouns and verbs to control the influence of language on learning, and there was no main effect of experimental language. Hebrew was used as the target language in Study 3, but out of concern for controlling differences between nouns and verbs I presented only a single inflectional example of each verb during learning and testing to control against inflectional differences between nouns and verbs. Evidence for a noun bias appeared in that study even so, which stands as evidence that the noun bias, as it seems to exist among adults, does so apart from greater morphosyntactic complexity of verbs than nouns. Although I have elsewhere pointed to the usefulness of the referential ambiguity hypothesis, a summary account of its success can be given as a way to convey its theoretic success in the context of this dissertation. Word imageability, goodness of depiction, and number of alternatives were significant predictors of name agreement which is really quite similar to saying these were significant negative predictors of referential ambiguity. Referential ambiguity was not highly reliable across models but closely approached significance in the one week model in Study 2 and produced a competing and significant model in Study 3. The fact that name agreement and word class rendered one another insignificant in Study 3 learning models was evidence that the noun bias might be at least partly attributable to noun-verb differences in name agreement; that is, perhaps the noun bias was due to greater referential ambiguity in verbs. The parts before relations hypothesis grew out of the referential ambiguity hypothesis; the support

found for the former may be taken as support for the latter. Finally in Study 3, poorer word learning under the one situation learning condition was that words and meanings were always presented with referential ambiguity!

Other possible predictors of word learning

In general, word learnability was primarily a product of word class and manipulated conditions. In Study 3 it was seen that cross-situational learning was much more efficient than one-situation learning as a word learning strategy. Study 2's inferential conditions may have acted negatively against nouns more than against verbs, but cross-situational learning acted positively on behalf of both word classes (relative to one-situation learning), probably due to presence of a second unique situational presentation of each target. Cross-situational learning might also have been superior to one-situation learning because as words were used again in a second unique situation, they came to be perceived by learners as more useful. A usage-base (or social-pragmatic) account of learning is one in which words that appear to be more useful are learned before words perceived to be less useful (Tomasello, Call, Behne, & Moll, 2005). Cross-situational learning is to be heralded for its successful elevation of word learning here.

Imageability and number of alternative interpretations were both decent predictors of name agreement, but only imageability was a reliable indicator of word learning, and only in Study 2 at five minutes. The absence of this effect at one week indicates this beneficial imageability effect is short-lived; the absence of this effect in Study 3 suggests this effect is not very robust. It seems, based on these results, unlikely that stable forms of word knowledge greatly depend upon how easily representations can be drawn to mind.

Delay was manipulated to test the hypothesis that words learned by inference would be less forgotten with time than words learned ostensively. Some support was found for this: the

deleterious effect of delay was less strong on words learned inferentially than ostensively. Recognition of words learned inferentially came a little close to chance, so replication of this finding would be helpful. Ideally, for measuring the possible difference in forgetting functions between methods of learning, ostensive and inferential learning should be on equal footing to begin with (at five minutes), and performance not close to chance at either testing delay. Still if one accepts the validity of this result, it is exciting to see that inferring word meanings may slow the naturally declining likelihood of their later retrieval. If this effect is multiplied over repeated inferences, this effect could become quite large and could potentially be a major methodological improvement in learning new vocabulary. The finding that method of learning interacted with delay is one of the most valuable findings in this dissertation. However its detection was in an over-sized model of word learning, and although this was a predicted effect, it was found as part of a larger exploratory study. A confirmatory study is therefore needed to replicate this finding, one in which the number of factors and tests is appropriate relative to the number of participants contributing to data.

Manipulating delay allowed exploring how other predictors varied over time. Far fewer effects were significant at one week compared to five minutes. At five minutes, a model was developed showing that two occurrences was strangely more hurtful to recognition compared to one; greater imageability was helpful; nouns were learned better than verbs; ostensive learning outdid inferential learning; and this last effect was qualified by word class such that although inferential learning was associated with lower overall learning, it hurt learning verbs less than nouns. Word class was the only effect that remained reliable at one week—though its effect direction reversed. This last observation is also of great significance, but must await confirmation by replication.

In the over-sized model of Study 2 describing learning at both delays, several curious interactions with delay were found that warrant follow-up work. Experimental language, number of lists learned prior, and participants' English background were not themselves significant but significantly interacted with delay. Targets learned in a list of words, when another list was learned prior, suffered greater decline in recognition over time than did the first list; this indicates the interference between lists appeared to intensify over time. The size of this interaction effect was matched by the interaction with delay and participants' first language: when English was known less prominently than another language, delay seemed to intensify a disadvantage to non-native English users. Finally the interaction effect of experimental language with delay was observed in which the likelihood to word recognition in one assigned language declined faster than it did in the other assigned language. Future researchers should work to better understand these effects or at least to be wary of them and take measures to either statistically control them or counterbalance their effects across more important manipulations.

As there were few known predictors at one week, there must be many unknown predictors of performance at such long delays. Besides word class, what else can predict whether or not a word will be recognized at one week? Future research is needed to identify these unknowns. That so little is known to affect learning at one week has major implications for the way language assessment and teaching success is approached. Immediate quizzing may mislead instructors to believe certain methods are ultimately more successful than others when this research shows no significant differences by method of learning in ultimate success as measured at one week. One good project following from this dissertation research would be to consider cross situational target recognition at one week, at 30 seconds, and at five minutes or some other intermediate delay to see whether cross situational learning, whose benefit was strong and

apparent at a 30-second delay, has a slower forgetting function relative to one-situation or ostensive learning.

Conclusions

Gentner (1982) found a noun advantage in the young children she studied in all six studied languages. Piccin and Waxman (2007) found a noun advantage in children and adults guessing word meanings from video using the Human Simulation Paradigm. Bornstein et al. (2004) also found a noun advantage among young children at almost every vocabulary range selected (from checklist-reported vocabularies ranging from 0-50, 51-100,101-200, and 201-500 words), and across all seven language they measured. Bornstein et al. proposed there are four plausible explanations which are normally intermingled, thus complicating the task of deriving a precise account for the noun bias among first language learners: morphology, saliency (utterance final), frequency, and pragmatics. The experimental evidence I have collected speaks to the last, pragmatics, and contraindicates the other three possibilities among adults. Frequency was less important than imageability, and was non-predictive once imageability was accounted for; saliency was not manipulated but verbs were always in the utterance-final position which is usually considered a highly salient position within the utterance; and morphology was controlled in Study 2, yet a noun bias was still observed at five minutes. Pragmatics informed my accurate prediction that cross-situational learning would lead to greater performance than one-situation learning.

This dissertation has come to a close, but my research on word learning does not end here. I believe that perceived usefulness is a fruitful and valuable area for research on acquisition of vocabulary, and foreign language more generally. The next word learning experiment I want to do is to test word usefulness by manipulating word use on a social dimension. That is, I will

manipulate the level of perceived usefulness by manipulating the number of speakers who utter a phrase: many speakers, few speakers, or one speaker (between subjects). I am predicting greater learning of words spoken by many speakers than by fewer speakers, controlling for number of repetitions. I have chosen to tackle word learning, but it is only one of several important aspects of language learning, my underlying research interest. Other important areas are learning pragmatic, grammatical, and gestural aspects of communication. I hope that this dissertation research contributes toward development and implementation of research-based language instruction techniques and programs aimed at improving communication across peoples and borders.

APPENDIX A

Study 1 & 2 (48 nouns, 48 verbs)				Study 3 (24 nouns, 24 verbs)	
alligator	hippo	to bathe	to pray	baby	to drink
angel	horse	to bungee jump	to read	elephant	to sit
apple	kangaroo	to clap	to rock climb	king	to type
armadillo	king	to cook	to run	penguin	to paint
astronaut	mail carrier	to cry	to shout	sailor	to listen
baby	monkey	to dig	to sing	turtle	to spin
bear	moose	to dribble	to sit	apple	to jump rope
bird	nurse	to drink	to skateboard	doctor	to smoke
boy	octopus	to eat	to ski	fish	to kiss
car	penguin	to fish	to sled	nurse	to mop
cat	pig	to golf	to sleep	princess	to iron
computer	police officer	to hatch	to smoke	robber	to shout
cow	princess	to hug	to sneeze	boy	to skateboard
deer	rabbit	to iron	to snort	cat	to sleep
doctor	refrigerator	to jump	to spin	duckling	to cry
dog	robber	to jump rope	to surf	officer	to write
dragon	sailor	to kayak	to swim	telephone	to fish
duckling	sheep	to kiss	to talk	alligator	to dribble
elephant	spider	to knit	to type	cow	to dig
fire fighter	strawberry	to laugh	to walk	monkey	to clap
fish	telephone	to mop	to wave	moose	to sled
flower	turtle	to paint	to whisper	rabbit	to laugh
frog	witch	to parachute	to wink	refrigerator	to run
hedgehog	zebra	to point	to write	spider	to parachute

Context images (Study 1 & 2)

alligator dribbling
angel hugging
apple jump roping
armadillo climbing
astronaut bungee jumping
baby drinking
bear bathing
bird reading
boy skateboarding
car talking
cat sleeping
computer surfing
cow digging
deer winking
doctor smoking
dog singing
dragon hatching
duckling crying
elephant sitting
fireman pointing
fish kissing
flower sneezing
frog swimming
hedgehog waving
hippo knitting
horse snorting
kangaroo skiing
king typing
mail carrier praying
monkey clapping
moose sledding
nurse mopping
octopus eating
officer writing
penguin painting
pig cooking
princess ironing
rabbit laughing
refrigerator running
robber shouting
sailor kayaking
sheep golfing
spider parachuting
strawberry walking
telephone fishing
turtle spinning
witch whispering
zebra jumping

APPENDIX B

		Studies 1 & 2 (96 words)		Study 3 (48 words)	
		Nouns	Verbs	Nouns	Verbs
Concepts					
N=26	Familiarity	6.45	6.49	6.56	6.52
N=26	Imageability	6.67	6.68	6.78	6.74
N=20	Frequency	3.43	3.87	3.56	3.83
Images					
N=19	Name-isolate	0.90	0.86	0.91	0.89
N=11	Name-context	0.90	0.77	0.91	0.81
N=20	Goodness	4.91	4.75	4.89	4.82
N=20	Alternatives	3.77	5.60	4.29	4.92
Audio					
	Utter length	0.90	0.95	0.64	0.75

APPENDIX C

Nouns	Image condition	
	Isolate	Context
alligator	0.90	0.69
angel	0.99	1.00
apple	0.95	0.98
armadillo	0.65	0.73
astronaut	0.90	0.90
baby	0.96	1.00
bear	0.95	1.00
bird	1.00	0.78
boy	0.80	0.97
car	0.92	1.00
cat	0.95	1.00
computer	0.95	0.83
cow	0.90	1.00
deer	0.89	0.99
doctor	0.95	1.00
dog	0.95	1.00
dragon	0.87	0.81
duckling	0.76	0.72
elephant	0.90	1.00
fire fighter	0.85	0.91
fish	0.95	1.00
flower	0.97	0.90
frog	0.95	1.00
hedgehog	0.73	0.53
hippo	0.80	0.90
horse	0.92	1.00
kangaroo	0.91	0.98
king	0.85	1.00
mail carrier	0.92	0.79
monkey	0.95	0.98
moose	0.74	0.74
nurse	1.00	1.00
octopus	0.90	1.00
penguin	0.93	0.99
pig	0.99	1.00
police officer	0.96	0.91
princess	0.95	0.82
rabbit	1.00	1.00
refrigerator	0.95	1.00
robber	0.95	0.78

Verbs	Image condition	
	Isolate	Context
to bathe	0.85	0.83
to bungee jump	0.78	0.34
to clap	1.00	0.88
to cook	0.91	0.90
to cry	1.00	0.89
to dig	0.96	0.93
to dribble	0.88	0.19
to drink	0.89	0.77
to eat	0.89	0.92
to fish	1.00	0.98
to golf	0.75	1.00
to hatch	0.57	0.97
to hug	1.00	0.90
to iron	0.86	0.90
to jump	0.95	0.75
to jump rope	0.76	0.84
to kayak	0.80	0.69
to kiss	0.97	0.92
to knit	0.79	0.83
to laugh	0.91	1.00
to mop	0.82	0.83
to paint	1.00	0.98
to parachute	0.77	0.55
to point	0.95	0.70
to pray	0.99	1.00
to read	0.99	0.90
to rock climb	0.94	0.78
to run	0.90	0.91
to shout	0.96	0.33
to sing	0.94	1.00
to sit	0.70	0.66
to skateboard	0.64	0.90
to ski	0.83	0.90
to sled	0.75	0.74
to sleep	1.00	0.99
to smoke	1.00	0.90
to sneeze	0.57	0.72
to snort	0.48	0.31
to spin	0.74	0.08
to surf	0.94	0.98

sailor	0.84	0.48	to swim	0.93	0.73
sheep	0.99	1.00	to talk	0.84	0.80
spider	0.95	0.92	to type	0.97	0.86
strawberry	0.90	0.70	to walk	0.99	0.80
telephone	0.90	1.00	to wave	0.87	0.60
turtle	0.95	0.98	to whisper	0.74	0.68
witch	0.84	1.00	to wink	0.89	0.98
zebra	0.91	1.00	to write	0.93	0.78

APPENDIX D

Interpreting odds ratios for categorical variables requires deciding (sometimes arbitrarily) on a reference value for a given variable. For example the reference value of word order was noun-verb. With the reference value established, the odds ratio can be understood as the ratio of the odds of outcome when the predictor is at its alternate value (in this example, the verb-noun order) over the odds of outcome when the predictor is at its reference value. In the odds metric, odds ratios may be flipped to describe the odds of the reference value from the perspective of the alternate value. Using the word order effect found in the over-sized model of word learning for Study 2 in Table 3.4 as a concrete example, the effect size of word order is 1.17. This means that this model predicts when words are ordered verb-noun, the odds of their recognition is 1.17 times greater than when they are ordered at their reference value, noun-verb.

Interpreting odds ratios for continuous variables is only a little more complex. The reference value is always defaulted at the bottom of the scale. Using imageability as a concrete example, imageability values lay on a 1 – 7 scale, thus 1 was its reference value. The odds ratio of imageability was also 1.17 coincidentally, but on a scale of 1 – 7, the change in odds from minimum to maximum imageability values this is a larger change in odds (i.e., effect) than the effect of word order (whose odds were also 1.17). The odds ratio can be understood as the rate of change in predicted odds along a variable's scale. Thus with every one unit increase in imageability (say from 6.00 to 7.00) there is an associated change in the odds of successful recognition by a factor of 1.17. This odds ratio applies across the entire spectrum of measured (and unmeasured) values, and is a description of the effect size associated with a change in imageability of one incremental unit on the measured scale. That is, the odds of recognition at 7.00 are 1.17 times greater than the odds at 6.00, and the odds at 6.00 are 1.17 times greater than

the odds at 5.00, etc. Improving imageability from 5.00 to 7.00, the model predicts, is associated with improvement from the odds of recognition at 5.00 (whatever that might be) by 1.17 x 1.17, or 1.37 times greater. Therefore although this odds ratio appears small, it is a hefty effect size when considering the improved odds of recognition along the entire spectrum of predictor values. In this example, the entire spectrum of imageability values was rather limited, but predicted recognition when imageability was at its maximum 7.00, compared to when it was at its minimum, 4.42 (so 7.00 is 2.58 units higher) is calculated as $1.17^{2.58}$, or 1.50 times greater predicted odds. Thus all other things being constant, "dog" (whose imageability was 7.00) was 1.5 times more likely to be recognized from its nonsense word cue than "hedgehog" (whose imageability was 4.42).

Understanding odds ratios with interactions is less straightforward than with main effects. When an interaction is significant, one should not interpret the involved main effect odds ratios by themselves because these values are displayed in output at their values when all other variables are held at their reference values (CRMportals, 2006), and they should be qualified this way. Word order and imageability could be deciphered simply because they were not part of any interaction factors. However the effect of method of learning was involved in at least one other interaction, so when describing the effect of method of learning, one must qualify this description by the level of the other variables it interacted with. Thus the effect of method of learning, as displayed in Table 3.4, was .49 for nouns (the reference value of the word class variable) only; for verbs the effect of method of learning was different, namely .49 x 1.32 (the interaction factor's odds ratio) = .65. In other words, the model predicts that when nouns are learned inferentially, they have .49 times lower odds of successful recognition than when they are learned ostensively, but for verbs, this negative effect of inferential learning is a little

milder—verbs learned inferentially are only .65 times less likely to be recognized than when they are learned ostensively.

It is possible to convert odds ratios into likelihoods by the formula: (odds ratio / 1 + odds ratio) = likelihood. An odds of 1.00 means no effect, so the likelihood of success under this would be (1/(1+1)=.50) exactly 50% when all other model factors are controlled. Applying this to the effect of method of learning, the model specified that when all other variables are controlled, nouns were (.49 / 1 + .49 = .329) about 33% likely to be recognized when learned inferentially, and verbs were (.65/1+.65 = .394) about 39% likely to be recognized when learned inferentially. Calculating likelihoods of success for reference values from the perspective of alternate values involves flipping odds ratios. The likelihood of successful recognition of a noun learned ostensively is ((1/.49)/1+(1/.49) or 2.04/1+2.04 = .671) about 67%, and the likelihood of successful recognition of a verb learned ostensively is ((1/.65)/1+(1/.65) or 1.54/1+1.54 = .606) about 61%. Notice that averaging likelihoods of noun recognition at ostensive (67%) and inferential (33%) results in 50% average likelihood (i.e., no effect), and the same is true of verbs learned ostensively (61%) and inferentially (39%), which average 50%; this math indicates that model odds values are provided as values when all other values, including the model's intercept, are controlled so that they may be ignored.

REFERENCES

Akhtar, N., Jipson, J., & Callanan, M. A. (2001). Learning words through overhearing. *Child Development, 72*(2), 416-430.

Akhtar, N. & Montague, L. (1999). Early lexical acquisition: the role of cross-situational learning. *First Language, 19*, 347-358.

Alishahi, A., Fazly, A., & Stevenson, S. (2008). Fast mapping in word recognition: What probabilities tell us. *Proceedings of the 12th Conference on Computational Natural Language.*

Appleman, I. B. & Mayzner, M. S. (1981). The letter-frequency effect and the generality of familiarity effects on perception. *Perception & Psychophysics, 30*(5), 436-446.

Atkinson, R. C. & Juola, J. F. (1971). Factors influencing speed and accuracy of word recognition. Paper presented at the *Fourth International Symposium on Attention and Performance*.

Au, T., Dapretto, M., & Song, Y. (1994). Input vs. constraint: Early word acquisition in Korean and English. *Journal of Memory and Language, 33*, 567-582.

Balota, D., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (A travel update). In M. Traxler, & M. Gernsbacher, Eds.), *Handbook of Psycholinguistics (2nd Ed.)* (Chapter 9). San Diego, CA: Academic Press.

Berger, D. E. (1993). Introduction to multiple regression. Retrieved from wise.cgu.edu.

Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing.* (pp.185-205). Cambridge, MA: MIT Press.

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers, 33*(1), 73-79.

Brown, G. D & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition, 15*(3), 208-216.

Brysbaert, M. (1996). Word frequency affects naming latency in Dutch when age of acquisition is controlled. *European Journal of Cognitive Psychology, 8*(2), 185-193.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15,* 17–29.

Center for Research on Languages – International Picture-Naming Project. Image retrieved on June 2, 2012 from http://crl.ucsd.edu/experiments/ipnp/

CRMportals Inc. (2006). Interaction terms vs. interaction effects in logistic and probit regression. Retrieved on 1/17/2012 from http://www.crmportals.com/crmnews.

Davidoff, J. & Masterson, J. (1996). The development of image naming: Differences between verbs and nouns. *Journal of Neurolinguistics, 9*(2), 69-83.

DeBleser, R. D. & Kauschke, C. (2003). Acquisition and loss of nouns and verbs: Parallel or divergent patterns? *Journal of Neurolinguistics, 16,*213-229.

Ellis, A. W. & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 515-523.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 145-154.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Technical report no. 257* No. BBN-R-4854

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek, & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs.* (pp. 544-564). New York, NY, US: Oxford University Press.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*, 135-176.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C (2006). Hard words. *Language learning and Development, 1*(1), 23-64.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology, 28* (1), 99-108.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*, 499-510.

Greenfield, P. M. & Smith, J. (1976). *The structure of communication in early language development*. New York: Academic Press.

Greenfield, P. M. & Alvarez, M. G. (1980) Exploiting nonverbal context to promote the acquisition of word-referent relations in a second language. *Hispanic Journal of Behavioral Sciences, 2*(1), 43-50.

Gopnik, A., & Choi, S. (1990). Do linguistic differences lead to cognitive differences? A

cross-linguistic study of semantic and cognitive development. *First Language, 10*(3), 199-215.

Gopnik, A., & Choi, S. (1995). Names, relational words, and cognitive development in English and Korean speakers: Nouns are not always learned before verbs. In M. Tomasello, & W. E. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs.* (pp. 63-80). Hillsdale, NJ: Lawrence Erlbaum Associate.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods & Measures, 1*(1), 77-89.

Hosmer, D. W., Jr., & Lemeshow, S. (2000). Applied logistic regression (2nd ed.). New York, NY: John Wiley.

Horst, J. S. & Samuelson, L. K. (2008). Fast mapping but poor retention among 24-month-old infants. *Infancy, 13*(2), 128-157. doi: 10.1080/15250000701795598

Humphreys, G. W., Riddoch, M., J. & Quinlan, P. T. (1988). Cascade processes in picture identification. *Cognitive Neuropsychology, 5*(1), 67-104. doi: 10.1080/02643298808252927

Imai, M., Li, L., Haryu, E., Okada, H., Hirsh-Pasek, K., Golinkoff, R. M., & Shigematsu, J. (2008). Novel noun and verb learning in Chinese-, English-, and Japanese-speaking children. *Child Development, 79*(4), 979-1000.

Jacobs, J. S. (2003). *Hebrew for dummies.* NewYork, NY: For Dummies.

Jaswal, V. K., & Markman, E. M. (2003). The relative strengths of indirect and direct word learning. *Developmental Psychology, 39*, 745-760.

Kauschke, C., Lee, H., & Pae, S. (2007): Similarities and variation in noun and verb acquisition: A crosslinguistic study of children learning German, Korean, and Turkish. *Language and Cognitive Processes, 22*(7), 1045-1072.

Kauschke, C. & von Frankenberg, J. (2008). The differential influence of lexical parameters on naming latencies in German. A study on noun and verb image naming. *Journal of Psycholinguistic Research, 37*, 243-257.

Kersten, A. W., Smith, L. B., & Yoshida, H. (2006). Influences of object knowledge on the acquisition of verbs in English and Japanese. In K. Hirsh-Pasek, & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs.* (Chpt. 19). New York, NY, US: Oxford University Press.

Knight, S. (1994). Dictionary: The tool of last resort in foreign language reading? A new perspective. *Modern Language Journal, 78*, 285-299.

Krashen, S. & Scarcella, R. (1981). On routines and patterns in language acquisition and performance. *Language Learning, 28*(2), 284-300.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411-433.

Lutz, K. A. & Lutz, R. J. (1978). Imagery-eliciting strategies: Review and implications of research. *Advances in Consumer Research, 5,* 611-620.

Markman, E. M. & Wachtel, G. F, (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20(2),* 121-157. doi:10.1016/j.physletb.2003.10.071

Masterson, J. & Druks, J. (1998). Description of a set of 164 nouns and 102 verbs matched for printed word frequency, familiarity and age-of-acquisition. *Journal of Neurolinguistics, 11*(4), 331-354. doi:10.1016/S0911-6044(98)00023-2

Naigles, L. R. (1990). Children use syntax to learn verb meanings. *Journal of Child Language, 17*(2), 357-374.

Nitsch, K. E. (1977). Structuring decontextualized forms of knowledge. Unpublished doctoral disseration, Vanderbilt University, Nashville, TN.

O'Hanlon, C. G. & Roberson, D. (2007). What constrains children's learning of novel shape terms. *Journal of Experimental Child Psychology, 97,* 138-148.

Piccin, T. B. & Waxman, S. R. (2007). Why nouns trump verbs in word learning: New evidence from children and adults in the Human Simulation Paradigm. *Language Learning and Development, 3*(4), 295-323.

Pinker, S. (1994). *The language instinct: How the mind creates language*. New York, NY: William Morrow & Company, Inc.

Sandhofer, C., & Smith, L. B. (2007). Learning adjectives in the real world: How learning nouns impedes learning adjectives. *Language Learning and Development, 3*(3), 233-267.

Sandhofer, C., Smith, L., Luo, J. (2000). Counting nouns and verbs in the input: differential frequencies, different kinds of learning? *Journal of Child Language, 27,* 561-585.

Shady, M., & Gerken, L. (1999). Grammatical and caregiver cues in early sentence comprehension. *Journal of Child Language, 26*(1), 163-175.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359-1366.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*(1-2), 39-91. doi: 10.1016/S0010-0277(96)00728-7.

Snedeker, J. & Gleitman, L. R. (2004). Why is it hard to label our concepts? In D. G. Hall and S. R. Waxman (Eds.), Weaving a lexicon (pp. 257-293). Cambridge, MA: MIT Press.

Stadthagen-Gonzalez, H., and Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavioral Research Methods, 38*(4), 598-605.

Székely, A., D'Amico, S., Devescovi, A., Federmeier, K., Herron, D., Iyer, G.,…Bates, E. (2003). Timed picture naming: Extended norms and validation against previous studies. *Behavior Research Methods, Instruments, & Computers 35* (4), 621-663.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental Psychology, 32*(3), 492-504.

Tardif, T., Shatz, M., & Naigles, L. (1997).Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. *Journal of Child Language, 24,* 535-565.

Tomasello, M. (1992). The social bases of language acquisition. *Social Development, 1*(1), 67-87.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*, 675 – 691.

Verspoor, M. & Lowrie, W. (2003). Making sense of polysemous words. *Language Learning, 53*(3), 547–586.

Vitevitch, M. S. & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40,* 374–408.

Vittinghoff, E. & McCulloch, C. E. (2006). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology, 165*(6), 710-718.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition, 109*, 163-7.

von Restorff, H. (1933). Über die wirkung von bereichsbildungen im spurenfeld. *Psychological research, 18*(1), 299-342.

Watkins, M. J. & Watkins, O. G. (1976). Cue-overload theory and the method of interpolated attributes. *Bulletin of the Psychonomic Society, 7*(3), 289-291.

Waxman, S. R., & Lidz, J. L. (2006). Early word learning. In D. Kuhn, R. S. Siegler, W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Vol 2, Cognition, Perception, and Language (6th ed.)*. (pp. 299-335). Hoboken, NJ: John Wiley & Sons Inc.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17,* 143-154.

Woodward, A. L., Markman, E. M. & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, *30*, 553-566.

Yeni-Komshian, G. H., Robbins, M., & Flege, J. E. (2001). Effects of word class differences on L2 pronunciation accuracy. *Applied Psycholinguistics, 22*(3), 283-299.

Yu, C. and Smith, L. B. (2007) Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414-420.