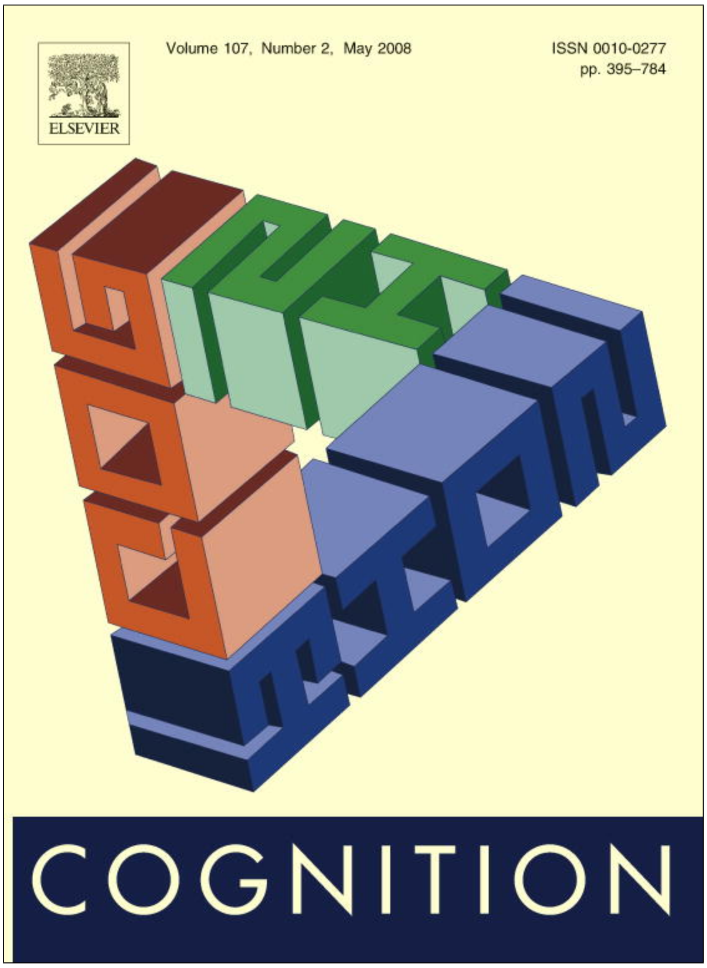
Volume 107, Number 2, May 2008
ISSN 0010-0277
pp. 395–784
ELSEVIER
COGNITION

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Cognition 107 (2008) 729–742
COGNITION
www.elsevier.com/locate/COGNIT

Brief article

Fine-grained sensitivity to statistical information in adult word learning

Athena Vouloumanos *

Department of Psychology, McGill University, Montréal, Qc, H3A 1B1, Canada

Received 20 September 2006; revised 22 August 2007; accepted 28 August 2007

Abstract

A language learner trying to acquire a new word must often sift through many potential relations between particular words and their possible meanings. In principle, statistical information about the distribution of those mappings could serve as one important source of data, but little is known about whether learners can in fact track multiple word–referent mappings, and, if they do, the precision with which they can represent those statistics. To test this, two experiments contrasted a pair of possibilities: that learners encode the fine-grained statistics of mappings in the input – both high- and low-frequency mappings – or, alternatively, that only high frequency mappings are represented. Participants were briefly trained on novel word–novel object pairs combined with varying frequencies: some objects were paired with one word, other objects with multiple words with differing frequencies (ranging from 10% to 80%). Results showed that participants were exquisitely sensitive to very small statistical differences in mappings. The second experiment showed that word learners' representation of low frequency mappings is modulated as a function of the variability in the environment. Implications for Mutual Exclusivity and Bayesian accounts of word learning are discussed.

Keywords: Language acquisition; Word learning; Statistical learning; Adults; Constraints

* Present address: Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA. Tel.: +1 212 998 7937; fax: +1 212 995 4866.
E-mail address: athena.vouloumanos@nyu.edu

0010-0277/$ - see front matter
doi:10.1016/j.cognition.2007.08.007

1. Introduction

As Quine (1960) noted, learning words is hard. A learner may see an orange-petalled object and hear the words "tulip", "wilting", and "spring". Multiple words are likely to co-occur with any given object: words with clearly relevant mappings (tulip), words with infrequent mappings that might be equally relevant (wilting), and other words whose occurrence in the context of that object may be purely accidental (spring). A word learner who observes the multiplicity of possible mappings between objects and words has a number of options. Such a learner may focus only on the most plausible or most frequent mapping, or, alternatively, track a broad range of mappings, in anticipation of further information. Moreover, to the extent that a learner tracks multiple mappings, that learner might or might not keep track of the frequency of occurrence of each type of mapping.

While is it well known that human learners can extract a variety of statistics from a wide range of stimuli (e.g., Creel, Newport, & Aslin, 2004; Fiser & Aslin, 2001; Gardner, 1957; Gómez, 2002; Gómez & Gerken, 1999; Hudson Kam & Newport, 2005; Kirkham, Slemmer, & Johnson, 2002; McDonald & Shillcock, 2003; Newport & Aslin, 2004; Reber, 1989; Saffran, Johnson, Aslin, & Newport, 1999; Saffran, Newport, & Aslin, 1996b; Turk-Browne, Junge, & Scholl, 2005; Maye, Werker, & Gerken, 2002), little is known about whether human learners track multiple referent–word relations and, if they can, whether learners are sensitive to the frequency of those mappings.

The present study asks whether adults track statistical information in the acquisition of word–object mappings, and examines the precision with which such statistical information is represented. To the extent that adults track mapping statistics, do they retain only information about high-frequency mappings, or do they keep track even of lower-frequency associations?

Given that words only sometimes co-occur with referents (Gleitman, 1990; Harris, Jones, & Grant, 1983), information about lower-frequency mappings could potentially be valuable, serving as a prerequisite to cross-situational learning (e.g., Roy & Pentland, 2002; Siskind, 1996; Yu & Smith, 2007), or allowing learners to entertain overlapping hypotheses about possible meanings of a word. Moreover, detailed statistical information about a range of mappings could allow learners to apply the constraint of mutual exclusivity (which assumes only one label per object) in a graded fashion (Regier, 1996) rather than as an all or none constraint (Markman, 1994).

Empirically, we might expect one of two patterns. Presented with a probabilistic environment, learners might use statistical structure to threshold between high and low probabilities, only encoding the highest available statistic (a possibility consistent with previous statistical language learning studies, e.g., Hudson Kam & Newport, 2005; Saffran, Aslin, & Newport, 1996a; Saffran et al., 1996b). Alternatively, acquisition processes might encode multiple statistics. Extant studies of word learning cannot distinguish between these two accounts.

The current studies contrast these possibilities by systematically varying word–referent co-occurrence statistics. Participants saw 12 novel objects paired with 12 novel words. The overall frequencies of each object and label were equal (10 times each), but the frequency of any given mapping varied between 1 and 10 co-occurrences (out of 10 total presentations; see Table 1). (We leave open the question of how learners might internally represent these statistics, e.g., in terms of probabilities or absolute frequencies, see Aslin, Saffran, & Newport, 1998 for discussion.) After exposure, learners' sensitivity to co-occurrence statistics was examined in two ways. (I) *Unambiguous* test trials – in which the test word had occurred with one test object either 10, 8, 6, 2 or 1 time(s) during training, but *never* occurred with the second test object – examined how learning of statistically variable correspondences is reflected in accuracy and reaction time. (II) *Ambiguous* test trials – in which *both* test objects had previously occurred with the test word, but with different probabilities – examined whether learners represent multiple word–object mappings in a fashion that retains information about their relative probabilities.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Forty undergraduate students at the University of British Columbia gave informed consent and received course credit for their participation.

2.1.2. Stimuli

2.1.2.1. Auditory stimuli. Twelve novel words were recorded by a native English-speaking female (SoundEdit Pro v.2, Macromedia, San Francisco, CA) and con-

Table 1
Novel word–novel object pairings used in Experiment 1

		Words											
		Det				vhP				hiP			
		W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12
Objects: Det	O1	10											
	O2		10										
	O3			10									
	O4				10								
Objects: vhP	O5					8			1	1			
	O6						8	1					1
	O7					1		8			1		
	O8						1		8			1	
Objects: hiP	O9							1		6	1	2	
	O10					1				1	6		2
	O11						1				2	6	1
	O12								1	2		1	6

Pairings were of three types: deterministic (Det), very high probability (vhP), and high probability (hiP).

sisted of consonant–vowel–consonant (C_1VC_2) syllables with consonants /p, t, s, n, k, d, g, b, m, l/ and vowels /æ, i, a, e, ^, u/. Each consonant was used once in C_1 position and once in C_2 position, and never repeated within a word. Place of articulation, a particularly salient feature for language-learners (Rice & Avery, 1995), was controlled both within and between words (e.g., the pattern $C_{(labial)}VC_{(coronal)}$ occurred only once). Phonotactic probabilities of positional segment frequency (e.g., of a particular consonant in word-initial position; average = .1431) and biphone frequency (i.e., of a particular C_1V or VC_2 combination; average = 0.00482) were within the range used in previous non-word studies (e.g., Vitevitch & Luce, 1999) (frequency counts by M. Vitevitch, personal communication February 8, 2001).

2.1.2.2. Visual stimuli. Twelve novel three-dimensional objects differing in colour and shape (e.g., Smith, Jones, & Landau, 1992) moved horizontally as a cohesive bounded unit (see Fig. 1; Strata 3D, St. George, UT; QuickTime, Apple, Cupertino, CA).

2.1.3. Design and procedure

The experiment consisted of a training phase in which novel word–novel object associations were presented, and a testing phase in which subjects were asked to pair the words with one of two objects. All participants were tested individually using a custom-scripted Hypercard stack (Apple, Cupertino, CA) on an Apple PowerMac G4 computer. Participants were tested on one of four different versions with different sound–object combinations. Training order and testing order were randomised for every participant.

2.1.3.1. Training. Participants were given no explicit instructions about the nature of the study, but were asked simply to pay attention to the words and objects. Participants were familiarised with a series of 120 word–object pairings during a 7 min training session (each object was presented for 3000 ms separated by a 500 ms ITI, with words playing 500 ms after trial onset). The 12 words and 12 objects in the training set were presented exactly 10 times each. Though objects (and words) in the deterministic group always co-occurred with exactly one word (object), objects (and words) in the probabilistic groups co-occurred with multiple words (objects) (see Table 1). For example, object O5 would co-occur 8 times with word W5, but 1 time with W7, and 1 time with W10. Thus, there were five levels of co-occurrence in the training phase: 10, 8, 6, 2, and 1.

2.1.3.2. Test. Testing was conducted using a two-alternative forced-choice design. Participants saw two objects moving simultaneously across the screen while one word was played, and selected the object that went "best" with the word by a button press. Two types of trials were tested: (I) In *unambiguous trials,* only one of the objects had co-occurred with the word during training, so there was only one correct answer (e.g., presenting O5 and O1 with W5 would be comparing 8 vs. 0 co-occurrences; an 8:0 trial). There were five such trial types: 1:0, 2:0, 6:0, 8:0, and 10:0. (II) In

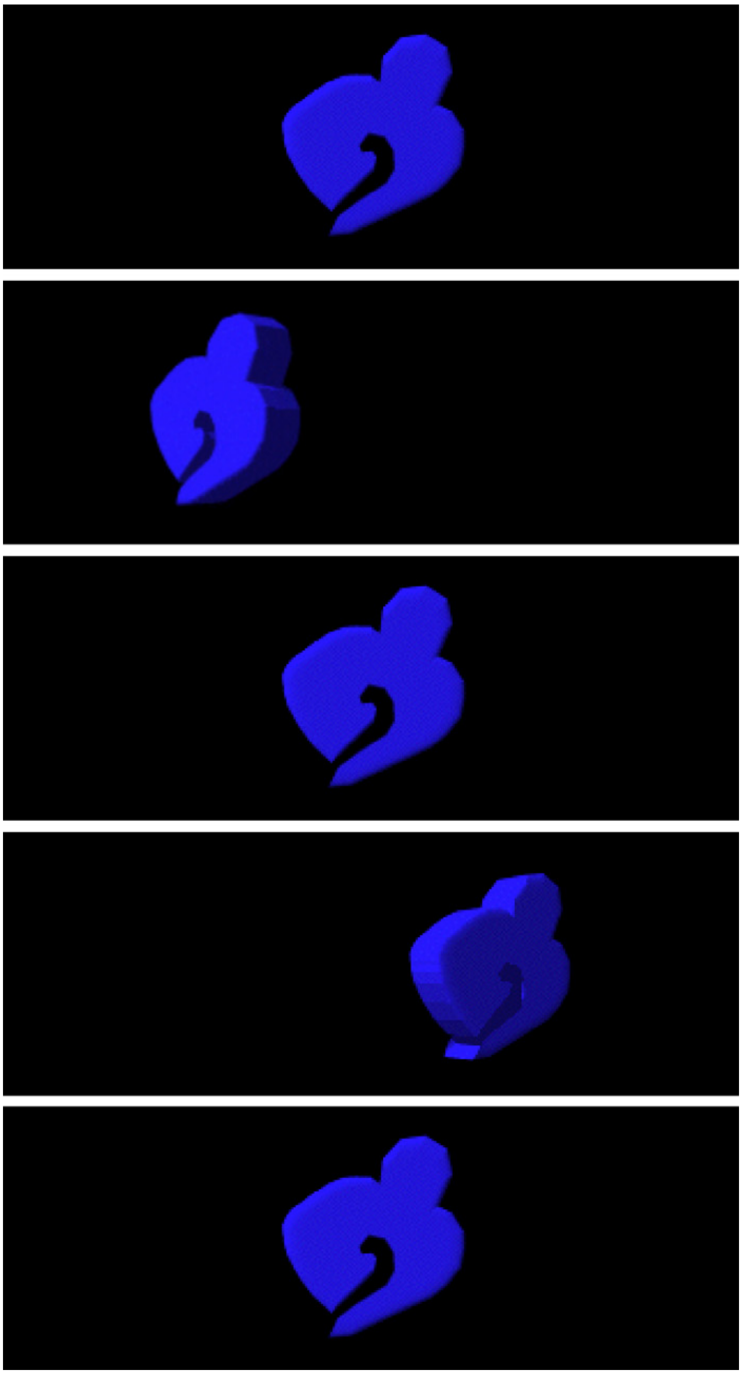

Fig. 1. An example of a novel 3D object used in this task. Five panels excerpted from the QuickTime movie show stills of the object as it moves in a horizontal arc across the screen, beginning and ending in a central position in the frame.

ambiguous trials, the word had co-occurred with *both* objects during the training phase, so that there were in fact two potentially correct answers, with one more frequent than the other (e.g., presenting O5 and O7 with W5 would be comparing 8 vs. 1 co-occurrences; an 8:1 trial). Four ambiguous trial types were presented: 2:1, 6:2, 6:1, and 8:1. Fifty-two test trials were played: six (1:0, 2:0) or eight trials (6:0, 8:0, and 10:0) for each unambiguous trial type, and four trials for each of the ambiguous trial types (2:1, 6:2, 6:1, and 8:1). Participants' accuracy and reaction times were recorded. A given pair of test objects was only presented once. The same objects or words were never presented on consecutive trials.

2.2. Results and discussion

2.2.1. Sensitivity to probabilistically occurring information

As there was no effect of version ($F < 1$), this factor was not included in subsequent analyses. Separate analyses of variance (ANOVAs) on unambiguous and ambiguous trials with test probability as a within-subjects factor (five levels for unambiguous: 1:0, 2:0, 6:0, 8:0, 10:0; four levels for ambiguous: 2:1, 6:2, 6:1, 8:1) revealed a significant effect of probability on accuracy in both unambiguous trials [$F(4, 156) = 31.82$, $p < .001$; as illustrated later, in Fig. 4 (diamonds), these data are captured by an exponential function], and ambiguous trials [$F(3, 117) = 7.37$, $p < .001$; Fig. 2]. To test whether ambiguous and unambiguous trials differed, we conducted an additional ANOVA on the 2:x, 6:x, and 8:x trials with trial type (unambiguous, ambiguous) and probability (2:x, 6:x, 8:x) as within-subject variables. Participants had significantly lower error rates on unambiguous trials ($M = .133$, $SE = .014$) than ambiguous trials [$M = .204$, $SE = .023$; $F(1, 72) = 7.52$, $p < .009$]. Performance at every probability ratio was better than chance. This exquisite performance is especially notable for the 2:1 condition in which participants chose between an object that had occurred twice with a word, and another object which had occurred only once, despite the fact that the word had occurred 6 times with still another (unavailable) object, and that each of the two objects had occurred 6 times with different (unavailable) words.

We next examined whether probabilities were reflected in reaction time (RT). We excluded individual trials for which RTs were beyond 2 standard deviations from that participant's mean. Analogous ANOVAs on RT data revealed a significant effect of probability on RT for both unambiguous [$F(4, 148) = 23.07$, $p < .0001$] and ambiguous trials [$F(3, 105) = 4.97$, $p = .003$], and these were marginally different from each other [$F(1, 32) = 2.85$, $p = .11$). Overall, both accuracy and RT measures present a probabilistic learning pattern.

2.2.2. Individual learning patterns

To ensure that the statistical learning pattern we observed was not an artefact of averaging across individuals (see Gallistel, Fairhurst, & Balsam, 2004 for group averaging effects on pigeon learning), we conducted an additional analysis on the 6:2:1:1 mappings, the only object–word pairing for which learners could have formed more than two statistical mappings. We calculated how many words were successfully mapped in both the 6:2 condition and the 2:1 condition, as compared with those only mapped in the 6:2 condition, the logic being that if the observed statistical learning

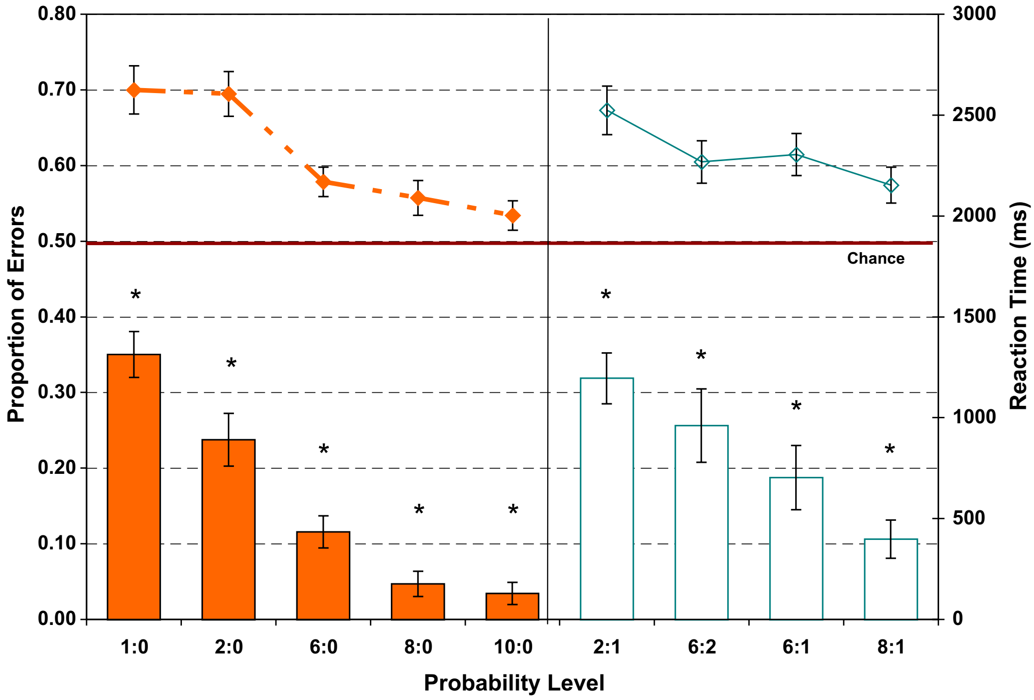

Fig. 2. Accuracy (bars, left y-axis) and reaction time (lines, right y-axis) in a probabilistic word learning task. $^{*}p < .05$.

pattern was generated across individuals (such that 6 out of 10 participants learned the "6", 2 of 10 learned the "2", etc.), the same individual would not succeed on both 6:2 *and* 2:1 trials. Each of the low probability words (W9–W12) was tested for each participant, yielding (4 words × 40 participants) 160 word trials. In 120 of 160 instances, succeeded in pairing the word with the "6" object in 6:2 pairings. Critically, among this subset of 120 word trials, words were also paired with the "2" object relative to the "1" object more often than chance (76/120 = .63, binomial test, $p = .005$). Furthermore, RT for these 2:1 trials ($M = 2567$ ms, $SE = 103$) was significantly slower than for the 6:2 trials ($M = 2052$ ms, $SE = 89$; paired samples 2-tailed t-test, $t(69) = 3.93$, $p < .001$), confirming that probabilities were reflected in the time to access the word–object representation. Together, these results reflect the workings of a statistically tuned learning mechanism within individuals.

3. Experiment 2

The results of Experiment 1 demonstrate that low probability or dispreferred mappings were tracked and encoded by learners, which speaks against a word learning model in which only high probability information is tracked. As a further test of word learners' sensitivity to precise details of the input, we investigated whether performance would be influenced by the variability of the environment. Suggestive evidence from Hudson's (2002) study on the acquisition of determiners (e.g., the, a) shows that learners' use of the dominant determiner form was sensitive to the distribution of the noise forms, such that learners were more likely to regularize the main form when the noise forms were more numerous. For example, if learners heard Determiner *Y* used 60% of the time, they would use it 60% of the time if there were fewer (e.g., 2) noise forms, but use it 85% of the time if there were more (e.g., 16) noise forms. Similarly, Gómez (2002) demonstrated that increased variability in the middle element in a*X*b language (e.g., an *X* set of 24 vs. only 3 elements) resulted in better learning of the perfectly deterministic a–b non-adjacent dependency. These studies suggest that the output of statistical learning can be affected by the statistical structure of the environment, with low variability leading to probability matching and high variability leading to rule-learning – in Hudson's study, the distribution of the low frequency form affected learning of the high frequency form, and in Gómez's study, the number of intervening forms influenced learning of a non-adjacent dependency – however, neither of these studies directly examined the effect of variability on the acquisition of low frequency forms[1]. Experiment 2 examines whether exposing learners to a more or less variable environment affects the representation of high and low frequency mappings differently.

3.1. Methods

3.1.1. Participants

Forty undergraduate students at the University of British Columbia gave informed consent and received course credit for participating in the study.

3.1.2. Stimuli

Identical to Experiment 1.

3.1.3. Design and procedure

Training and test were conducted as in Experiment 1, however the structure of the environment differed. Instead of word–object pairings ranging from 1–10, the maximum likelihood of word–object co-occurrence was 6 (see Table 2). Only three training probabilities were presented (6, 2, and 1), creating a noisier environment because very high probability and deterministic pairs were absent. Testing was conducted using a two-alternative forced-choice design. As in Experiment 1, there were two trial types of interest: (i) *unambiguous trials* (1:0, 2:0, 6:0), and (ii) *ambiguous trials* (2:1, 6:2, 6:1). Eight test trials of each probability condition were presented. In addition, we added a small number of trials (four) for which there was no correct answer (0:0 trials) as a check against bias in the pairings, for which performance was at chance, as expected. Object pairings and presentation were randomized as in Experiment 1.

[1] One of Hudson's analyses (Fig. 3) calculated difference scores between participants' ratings of the main determiner form and the noise forms, but did not examine learning of the noise forms independently of the main forms and thus does not speak directly to the issue at hand.

Table 2
Novel word–novel object pairings that constitute the more variable environment of Experiment 2

			Words											
			hiP											
			W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12
Objects	hiP	O1	6				1				1			2
		O2	2	6				1		1				
		O3		2	6				1				1	
		O4			2	6				1		1		
		O5		1		2	6				1			
		O6			1		2	6				1		
		O7				1		2	6					1
		O8	1						2	6			1	
		O9			1			1		2	6			
		O10							1		2	6		1
		O11	1			1						2	6	
		O12		1			1						2	6

3.2. Results and discussion

3.2.1. Sensitivity to statistically variable information

We examined whether learners encoded the training probabilities by conducting separate ANOVAs on unambiguous and ambiguous test trials with probability level as a within-subjects factor (unambiguous: 1:0. 2:0, 6:0; ambiguous: 2:1, 6:2, 6:1). As in Experiment 1, there was a significant effect of probability on error rate in both unambiguous trials, $F(2,78) = 19.27$, $p < .001$, and ambiguous trials, $F(2,78) = 14.84$, $p < .001$. Performance in all probability conditions was better than chance (see Fig. 3). As in Experiment 1, a 2 (probability level: 2:x, 6:x) by 2 (unambiguous vs. ambiguous) ANOVA yielded significantly lower error rates in unambiguous ($M = .295$, $SE = .021$) as compared to ambiguous trials [$M = .341$, $SE = .016$), $F(1,39) = 4.46$, $p < .05$]. Notably, error rates for all probability levels of Experiment 2 differed from chance indicating that participants had still encoded some low probability information. Interestingly, as illustrated in Fig. 4 for both Experiments, an exponential function captures the relationship between accuracy and frequency, in keeping with much of the literature on human memory (Anderson, 1995).

Analogous ANOVAs on RT also yielded a significant effect of probability for unambiguous trials, $F(2,78) = 9.15$, $p < .001$, and a significant effect for ambiguous trials, $F(2,78) = 4.16$, $p < .05$. There was no effect of ambiguity on RT, $F(1,39) < 1$, *ns*.

3.2.2. Effect of the statistical structure of the environment on statistical representations

To determine the effect of environmental variability on the representation of high and low probabilities, we compared participants' performance in the less variable environment of Experiment 1 with the more variable environment of Experiment

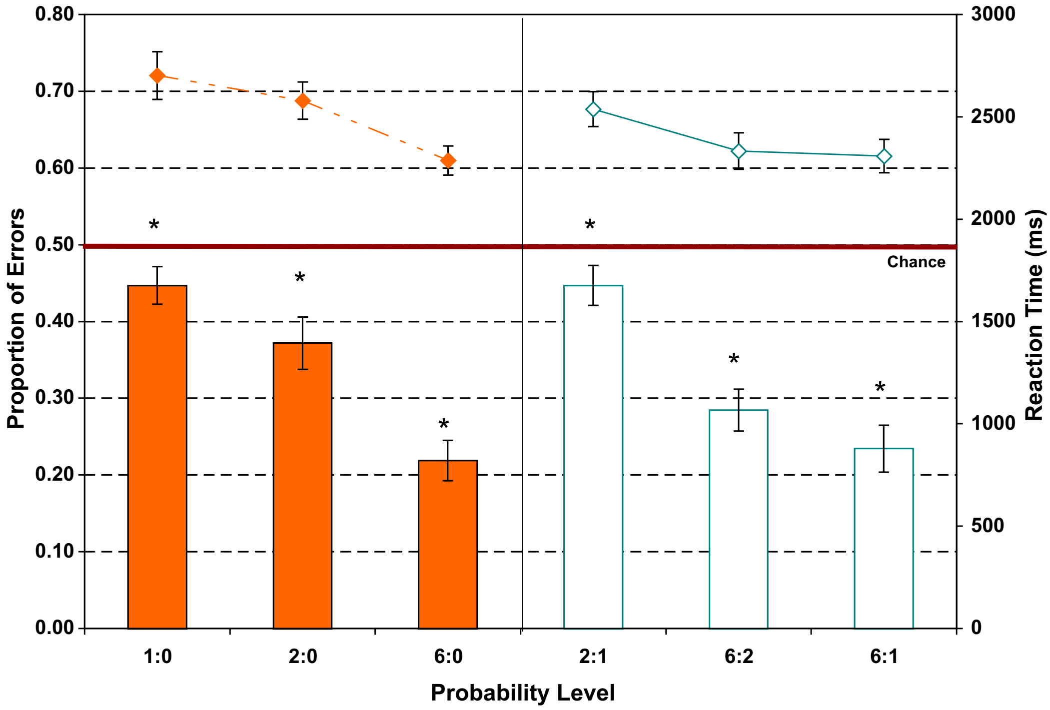

Fig. 3. Accuracy (bars, left *y*-axis) and reaction time (lines, right *y*-axis) for probabilistic word learning in a more variable environment. $^{*}p < .05$.

2. In light of previous findings (e.g., Gardner, 1957; Gómez, 2002; Hudson, 2002), we predicted that the more variable environment of Experiment 2 would affect learning of high (6) and low (2, 1) probabilities differently. To test this hypothesis we conducted planned comparisons at every probability level that was comparable between the two experiments: the 2:1, 6:2, and 6:1 test trials[2]. Performance on 2:1 trials was significantly worse in the more variable group [$F(1,78) = 9.55$, $p = .003$], while performance for both 6:2 [$F(1,78) < 1$, *ns*] and 6:1 [$F(1,78) < 1$, *ns*] trials was equivalent between the groups (see Fig. 5). Increasing the variability of the learning context thus modulates learning of low probability mappings.

4. General discussion

Successful statistical learning is often implicitly equated with noticing the alternative that is most frequent or of greatest cue validity. In such a framework, the

[2] The unambiguous test trials were not compared because the "0" foils were different in the two experiments. This was necessary to balance the number of times each object appeared as correct and incorrect in the test trials. Specifically, in Experiment 1, the foils were chosen from the deterministic and very high probability pairs, while in Experiment 2, foils were all high probability pairs. The ambiguous trials were thus the only trials comparable across both experiments.

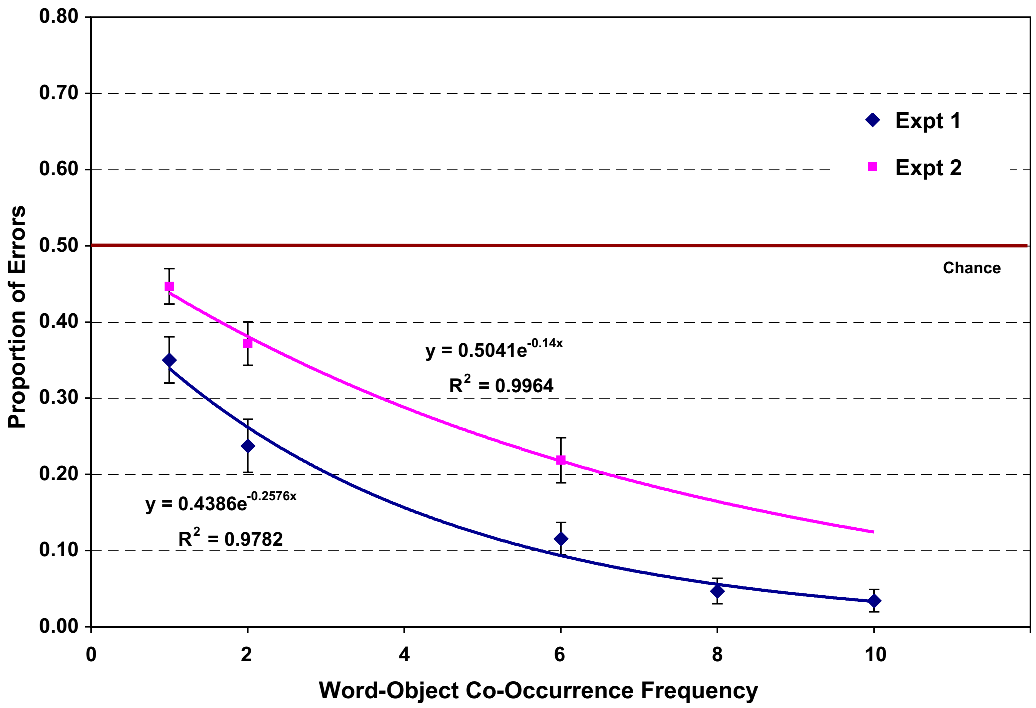

Fig. 4. Data from unambiguous trials of Experiments 1 (diamonds) and 2 (squares) represented on a linear scale (with exponential curves superimposed).

low frequency cases would fade from view, untracked. Instead, this study found that adult learners were exquisitely sensitive to co-occurrence statistics between words and objects, and even differentiated between the probabilities of infrequent (dispreferred) mappings, treating a pairing that occurred 20% of the time as more likely than a 10% pairing. These results extend the conclusions of recent studies demonstrating the robustness of statistical learning in humans (e.g., Fiser & Aslin, 2001; Maye et al., 2002; Newport & Aslin, 2004; Saffran et al., 1996a) to the domain of word learning, but also go beyond them, by establishing that learners can represent precise information about a range of possible word–referent mappings, even dispreferred ones. A sensitivity to a range of mapping probabilities might be crucial in word learning, allowing the learner to keep track of a number of candidates, and not just mappings with the highest co-occurrence statistics. Significantly, this suggests that learning a word–referent mapping is not necessarily an all-or-none process. In particular, putative word learning constraints like mutual exclusivity (Markman, 1994) might be more plausibly applied in a graded fashion (Regier, 1996). For example, learners could entertain overlapping hypotheses about the referents of a word, and assign different likelihoods to each of these candidate mappings.

In real-world acquisition situations, it might not be feasible to keep track of all available statistical information. Memory and attention considerations constrain the products of statistical learning (e.g., Kareev, 1995; Turk-Browne et al., 2005).

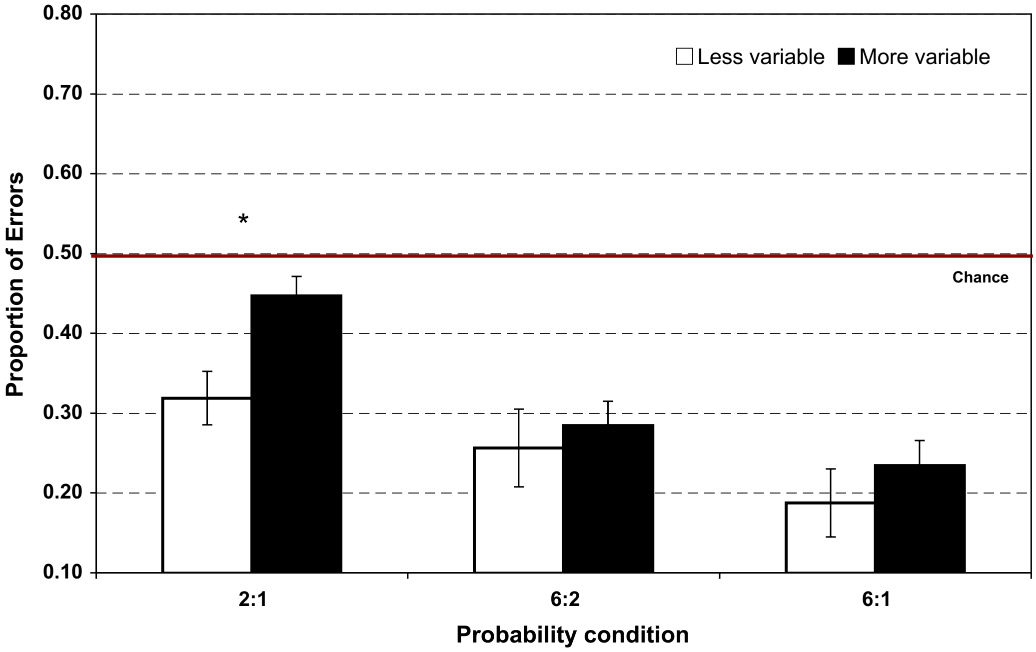

Fig. 5. The effect of environmental context: Participants' accuracy in a less variable (open squares) versus a more variable environment (filled squares).

Here, we found that the variability of the environment is another factor that modulates the outcome of statistical analysis. Exposing adult learners to the more variable environment of Experiment 2 constrained statistical learning such that learning of lower frequency mappings was reduced relative to high frequency mappings.

Detailed statistical representations have been demonstrated to play a role in other human learning systems, for example, in sensorimotor learning, in which representations of prior probabilities and sensory feedback interact (Hunt & Aslin, 2001; Kording & Wolpert, 2004), and might be a hallmark of learning systems that operate according to Bayesian principles (e.g., Gopnik et al., 2004). As Bloom (2000) has argued, word learning requires more than simple association, one attractive possibility is that word learning may depend on a Bayesian merger between sophisticated hypothesis testing and rational statistical inference (Xu & Tenenbaum, 2005, 2007). The current paper speaks to the viability of one important aspect of such a proposal, inasmuch as any Bayesian system would require that learners posses a "graded sensitivity to uncertainty in prior knowledge and the input" (Xu & Tenenbaum, 2005). The present data show that this particular requirement is clearly met: learners can, with great precision, track statistically variable mappings between words and their referents. Such precisely tracked mappings may then serve as a foundation for the more complex computations required in a complete system for word learning.

Acknowledgements

This work was supported by funding from the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Fonds québécois de la recherche sur la société et la culture to the author, and funding from NSERC and the Human Frontier Science Program to Janet Werker. I am especially grateful to Janet Werker without whom this work would not have been possible, and Gary Marcus and Kris Onishi for crucial discussions and comments on previous drafts. My thanks to Katherine Yoshida and Miguel Imperial, as well as Karla Kaun, Alisa Almas, Marisa Cruickshank, Michael Vitevitch, Rhonda Amsel, Jesse Snedeker, and Alia Martin for their contributions to this research.

References

Anderson, J. R. (1995). *Learning and memory: An integrated approach*. New York: Wiley.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*(4), 321–324.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of non-adjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(5), 1119–1130.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*(6), 499–504.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America, 101*(36), 13124–13131.

Gardner, R. A. (1957). Probability-learning with two and three choices. *American Journal of Psychology, 70*(2), 174–185.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition, 1*(1), 3–55.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431–436.

Gómez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*(2), 109–135.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and bayes nets. *Psychological Review, 111*(1), 3–32.

Harris, M., Jones, D., & Grant, J. (1983). The non-verbal context of mothers' speech to infants. *First Language, 4*(10), 21–30.

Hudson, C. L. (2002). *Pidgins, creoles, and learners: How children and adults contribute to the formation of creole languages*. Rochester, NY: University of Rochester.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Leaning and Development, 1*(2), 151–195.

Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General, 130*(4), 658–680.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition, 56*(3), 263–269.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*(2), B35–B42.

Kording, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*(6971), 244–247.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua, 92*, 1–4.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14*(6), 648–652.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127–162.

Quine, W. V. (1960). *Word and object*. Cambridge, MA: Technology Press of the Massachusetts Institute of Technology.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118*(3), 219–235.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. Cambridge, Mass: MIT Press.

Rice, K., & Avery, P. (1995). Variability in a deterministic model of language acquisition: A theory of segmental elaboration. In J. Archibald (Ed.), *Phonological acquisition and phonological theory* (pp. 23–42). Lawrence Erlbaum Associates, Inc.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26*(1), 113–146.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*(4), 606–621.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*(1–2), 39–91.

Smith, L. B., Jones, S. S., & Landau, B. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. *Developmental Psychology, 28*(2), 273–286.

Turk-Browne, N. B., Junge, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552–564.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*(3), 374–408.

Xu, F., & Tenenbaum, J. B. (2005). Word learning as Bayesian inference: Evidence from preschoolers. Paper presented at the Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society, Stresa, Italy.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245–272.

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