Leveraging mutual exclusivity for faster cross-situational word learning: A theoretical analysis

Shohei Hidaka (shhidaka@jaist.ac.jp) and **Takuma Torii (tak.torii@jaist.ac.jp)**
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan

George Kachergis (george.kachergis@gmail.com)
Dept. of Artificial Intelligence / Donders Institute, Radboud University
Nijmegen, the Netherlands

Abstract

Past mechanistic accounts of children's word learning claim that a simple type of cross-situational learning is powerful enough to match observed rates of learning, even in quite ambiguous situations. However, a limitation in some of these analyses is their reliance on an unrealistic assumption that the learner only hears a word in situations containing the intended referent. This study analyzed a more general type of cross-situational learning based on the relative frequency of word-object pairs, and found it to be slower than the simple mechanism analyzed in prior work. We then analytically explored whether relative-frequency learning can be improved by incorporating the mutual exclusivity (ME) principle–an assumption that words map to objects 1-to-1. Our analyses show that with a certain type of correlation in word-to-word relationship, ME makes relative frequency learning as efficient as fast-mapping, which can learn a word in one exposure.

Keywords: Word learning; Cross-situational learning models; Mutual exclusivity; Language acquisition

Introduction

To a new learner of a language with a completely unknown word-referent mapping system, determining which words refer to which referents in any given scene may seem impossible on the face of it, since a word could refer not only to an object (e.g., 'apple'), but to a class of objects (e.g., 'fruit'), a property ('red'), or any one of endless possible combinations or configurations of features in the scene–an unconstrained problem of logical induction (Quine, 1960). In contrast to this theoretical observation about referential uncertainty, children are thought of as efficient learners, and in fact most human children do learn to understand and use an impressive number of words within the first years of life, achieving a vocabulary of roughly 60,000 by 18 years of age (Bloom, 2000). Developmental researchers have theorized that children use a variety of lexical constraints to limit the number of possible mappings they consider, and a number of empirical studies support these claims (Clark, 1987; Markman, 1990, 1992; Golinkoff, Hirsh-Pasek, Bailey, & Wegner, 1992). One lexical constraint, used here as in past theoretical accounts–and supported by empirical developmental data, is that learners are biased to map words to entire objects, rather than to a feature of an object, or a group/configuration of objects (Markman, 1990).

Beyond lexical constraints that reduce the number of hypothesis meanings considered for a given word in a given situation, another possible remedy for the contradiction between the difficulty of the unconstrained inductive account of word learning and the ease of the observed process is that learners also reduce uncertainty in the word-object map by statistical inference over time, based on observing word-object pairs across multiple situations. Cross-situational learning (Pinker, 1984; Akhtar & Montague, 1999; Siskind, 1996) is a type of learning based on this idea, which has been analyzed both empirically and theoretically over decades (Yu, 2008; Blythe, Smith, & Smith, 2010). Blythe et al. (2010) formally quantified the effect of a type of cross-situational learning in terms of the rate of vocabulary growth. More recent studies (Blythe, Smith, & Smith, 2016; Vogt, 2012) further showed that this type of cross-situational learning can be considerably slowed down for certain types of word co-occurrence distributions, including power-law distributions in which most words are seen relatively rarely, which describe word frequency distribution in natural languages (Zipf, 1949).

These theoretical analyses are still quite limited in their generality. The class of cross-situational learning analyzed in these past studies is called *eliminative learning.* In this scheme, when a learner is exposed to a set of referents, a correct word is spoken–never is a word spoken when its intended referent is not present. In this case, the learner can safely "eliminate" the possibility of word A being associated to the object B, if he or she experiences one episode that the word A is spoken without the object B. As this special assumption does not generally hold in real-world learning, the estimates on the speed of cross-situational learning in past studies give only an optimistic upper bound for its learning efficiency.

In this study, we consider a more general type of cross-situational learning, called *relative frequency* learning, of which eliminative learning is a special case. In the relative frequency learning scheme, it is assumed that a language system encodes the word-object pair with frequency higher than the other candidate pairs as the correct one, and the learner infers such relatively more frequent word-object pairs from the sample. Under this assumption, the eliminative learning scheme is identified with the special case of seeing the correct word-object pair with probability 1. In general, however, the eliminative learning rule cannot apply (or will mislead the learner if it is forced to apply) in word learning of a relative-frequency language system.

Therefore, relative frequency learning is generally slower than the eliminative learning. Thus, the main problem considered in this study is what plausible factor might make this type of learning more efficient – and can it be made efficient enough to be a realistic account for children's word learning? Specifically, we analyze the beneficial effect of learners applying a general principle of mutual exclusivity (ME), an assumption of a word-object regularity requiring that no two objects are associated to one word. Application of a ME principle has long been theorized to be a constraint that can speed children's word learning (Markman & Wachtel, 1988), and has found empirical support in both children (Halberda, 2003) and adults (Yurovsky & Yu, 2008; Kachergis, Yu, & Shiffrin, 2012). We then consider a word-word statistical relationship in which a group of distractor objects tend to co-occur with a word and thus slow learning.

In the following, we first outline the theoretical framework in which we provide a series of analyses of relative frequency learning. Second, we evaluate the basic learning efficiency in this scheme. Then we extend this evaluation of learning efficiency to multiple scenarios with different word-to-word statistical relationships.

Relative-frequency learning

Basic framework

In this study, we consider the following word learning scenario. The learner is exposed to multiple words and objects in each situation. In each situation, the learner does not know which word refers to which object, and the correct word-object mapping can only be inferred by integrating evidence across observations of multiple situations. Let $W = \{1, \dots, n\}$ be a set of words and $O = \{1, \dots, m\}$ a set of objects (or referents) which appear in these situations. In this study, we consider a language structure with one-to-one word-object mapping, in which $n = m$ and the word i refers to object i. This is a quite strong assumption, which may not be considered entirely realistic as it is. It offers, however, a first approximation upon which we can base the analysis and later extend it.

Here, we consider a particular word learning scheme, called *relative frequency learning*, in which each object's to-be-associated (i.e., 'correct') word is spoken in its presence with greater frequency than any other word. This is a code in the sense of information theory – the signal, the correct word-object mapping, is encoded in the statistical regularity in observation across situations (channel), and the learner decodes (infers) the correct word-object map using the underlying regularity: the correct word-object pair is the most frequent among the others.

There are theoretical analyses of a special case of this relative frequency learning, in which the correct word is spoken only in the presence of the corresponding object (i.e., $p(object|word) = 1$). In this special case, the learner can use not only the knowledge that the correct pair is more frequent, but also the quite strong rule that any object which does not appear with a spoken word cannot be the intended referent of that word. Thus, this learning scheme, which eliminates any word-object pair with probability less than 1 is called *eliminative learning* (Blythe et al., 2010). In this study, beyond this special case, we analyze a more general case of language and learning coded on the basis of relative frequency.

Formulation

Denote the frequency of object o given word w by $f(o \mid w)$. Then, suppose the learner (decoder) declares that the object $o \in O$ is the referent of the word $w \in W$ with probability

$$P(o \mid w) = \frac{e^{f(o|w)}}{\sum_{o \in \{O\}} e^{f(o|w)}}.$$

In this scheme, the error, wrong declaration of the correct object, for word w with the number of observed situations n is proportional to $\epsilon(n, w) := \sum_{o \neq w} e^{f(o|w) - f(w|w)}$. The sum of the errors for all words $\epsilon(n, w) := \sum_{w \in W} \epsilon(n, w)$ is an exponential function of the number of situations. Let us denote the rate of the exponential function as R, and thus $\epsilon(n) = e^{-Rn}$. For a code with rate R encoding less than e^{Rn} signals, $\lim_n \sum_{w \in W} P(o \mid w) = 1$, and thus it is said to be learnable (reachable in information-theoretic terminology). If the rate satisfies $\epsilon(n) = e^{-Rn} < e^{-Cn}$ for any code, the constant C is said to be the *capacity* of this channel in information-theoretic terms (Shannon, 1948). The rate, or the exponent coefficient of the error function, is a fundamental characteristic of the language-learning system when viewed as a signal transmitting process.

Efficiency

In the relative frequency learning scheme, the object o with the second largest probability given the word w, $p_{w|w} > p_{o|w} > p_{o'|w}$ for $o' \neq o, w$, is a key parameter giving the asymptotic time to learn the word w. With objects with the largest and second largest probability, the sample frequency can be written as follows. Let $\overline{p} = 1 - p$. Specifically, consider that the sample frequency $f_{now} = f_n(o|w)$ follows the binomial distribution

$$P(f_{now}|n, p_{ow}) = \binom{n}{f_{now}} p_{ow}^{f_{now}} \cdot \overline{p}_{ow}^{n - f_{now}}$$

with probability p_{ow}.

Given this, the error probability in learning is characterized as follows. The probability for the word w to be associated with the object o is proportional to $e^{f_{now}}$. For a sufficiently large n, the difference between the two random variables asymptotically approaches

$$\lim_{n\to\infty} \frac{e^{f_{now}-f_{no'w}}}{e^{n\Delta_{o,o'|w}}} = C,$$

where $\Delta_{o,o'|w} := \frac{p_{ow}-p_{o'w}}{p_{ow}\bar{p}_{ow}+p_{o'w}\bar{p}_{o'w}}$. If there are m objects with the second largest probability $p_{ow} > q > \max_{o'\neq w}, p_{o'w}$ for the word w, the error probability is $1 - P(w|w) \to Cme^{-n\frac{p_{ow}-p_{o'w}}{p_{ow}\bar{p}_{ow}+p_{o'w}\bar{p}_{o'w}}}$. Thus, the rate of the relative-frequency code is $R = \min_w \Delta_{w|w}$ where

$$\Delta_{o|w} := \frac{p_{ow} - \max_{o'\neq o} p_{o'w}}{p_{ow}\overline{p_{ow}} + \max_{o'\neq o} p_{o'w}\overline{\max_{o'\neq o} p_{o'w}}}.$$

This analysis implies that the word-object pair with the smallest margin to second largest probability decides the learning rate in the relative frequency code.

Incorporating mutual exclusivity (ME)

In the above analysis of the relative frequency code, the lexical constraint of one-to-one word-object mapping is not taken into consideration in the learning process. However, if the learner exploits the fact that no two objects are associated with the same word, namely correct word-object pairs are mutually exclusive, the learning is expected to be more efficient than the alternative without the knowledge. Let us call this ME learning. The difference in the rate of learning assuming ME and general relative frequency would be the effect of introducing a ME constraint in cross-situational learning.

With ME, the learner can exclude object o when learning word w, if the object o is likely to be associated with some other word $w' \neq w$. Thus, the learning order of the words has considerable impact in learning under ME. As the previous analysis shows that the second most probable objects for word w is the key factor giving the learning rate, let us call them *distractors against the word* w, and denote the set of distractors by $D(w) := \{w' | \max_{o\neq w} f_{o|w} = f_{w'|w}\}$.

Best- and worst-case scenarios

Here let us analyze ME learning under a simplification that the learning time for the words with no distractor is T_0 and that for the words with one more distractors is T_1. The former case with no distractor is said to be *fast mapping*, in which a particular word-object pair is presented alone in a situation, and the learner learns the pair in a single shot (Carey & Bartlett, 1978). The latter case is analyzed in the previous section in case of the relative frequency learning. In this case, if all the distractors has been eliminated, by the effect of ME, the corresponding object can be uniquely identified, which is effectively the same as fast mapping. Thus, the worst-case learning time approaches that of relative frequency learning, and the best-case learning time approaches that of fast mapping, as the number of words is sufficiently large.

Randomly distributed distractors

Random learning order Consider the case that each word is learned in a serial order and each has k distractors. Furthermore suppose that the learning order is a random permutation, namely any order is uniformly sampled. Figure 1 shows a schematic co-occurrence matrix of the five such word-object pairs (filled markers) with $k = 2$ randomly distributed distractors (open markers) for each pair. In this case, one expects that one word is likely to be learned after the k distractors with probability $1/(k + 1)$. This is exactly true, if the number of words n approaches to infinitely large. Therefore, the sum of expected learning time for all the words is

$$T = n\left(\frac{k}{k+1}T_1 + \frac{1}{k+1}T_0\right). \qquad (1)$$

Thus, when the learning order is a random permutation, the expected learning time is only the factor of $\frac{1}{k+1}$ shorter than the original time nT_1 at shortest.

Word	Objects					$\#D$
"Circle"	●	△		☆		2
"Triangle"		▲	□		◇	2
"Square"	○	△	■			2
"Star"	○			★	◇	2
"Diamond"		△	□		◆	2

Figure 1: A schematic word-object co-occurrence matrix in the case with random learning order and randomly distributed distractors.

Shared distractors

Best and worst learning order Let us consider the best and worst case by manipulating which words the k distractors are associated. In one of the best cases, every word shares the same set D of k distractors. Figure 2 shows a schematic co-occurrence matrix of the five such word-object pairs (filled markers), and each pair has $k = 2$ distractors (open markers) and most of words share the same two distractors. In this case, the shortest learning time is obtained by a sequence of learned words in which the k words with the k distractors as their correct objects first (required about T_1 time each) and the others later (required T_0 time each). In the example (Figure 2), one of the best order is to learn the word "Circle" and "Triangle" at the first two rows in the matrix, and then learn the other words. In this case, the total learning time is

$$T = kT_1 + (n-k)T_0.$$

As the number of words n gets larger with a constant k, the learning time approaches to that of the fast mapping (T_0 per word), which is the lower bound of learning time.

In one of the worst cases, on the other hand, the longest learning time is obtained by the reversed sequence, in which the words with the k distractors as their correct objects are learned last. In total, the longest learning time is

$$T = nT_1.$$

As the number of words n gets larger with a constant k, learning time approaches that of relative frequency learning, which is the upper bound of learning time.

Random learning order Thus, this analysis with the best and worst case scenario suggests that the learning order of words has a large impact on learning time. However, the expected learning time with the shared distractors is, again, exactly $1/(k+1)$, which is no better than the learning time of the case with k random distractors (Equation (1)):

$$T = n\left(\frac{k}{k+1}T_1 + \frac{1}{k+1}T_0\right).$$

This analysis suggests that even systematically shared distractors *cannot* improve the learning time on average, if the learning order is uniformly at random.

Word	Objects					$\#D$
"Circle"	●	△	□			2
"Triangle"	○	▲	□			2
"Square"	○	△	■			2
"Star"	○	△		★		2
"Diamond"	○	△			◆	2

Figure 2: A schematic word-object co-occurrence matrix in the case with random learning order and $k = 2$ distractors shared by all the words systematically.

Correlation in word-to-word relationship

Mixture of two groups of words

As the previous analysis suggests that the relative frequency learning of a one-to-one word-object map in the cross-situational setting is as slow as independent learning even by incorporating ME. This result is largely due to the statistical structure of the word-word relationship – in the previous analysis, each word has k other random words as distractors. In this section, we consider a specific class of statistical regularity in the word-word relationship. Specifically, suppose there are two groups of words: in the one group of words, each word has no distractor, and in the other group of words, each word has k distractors, whose referring words have no distractor (Figure 3). Thus, the learner is exposed to a mixture of two groups of words with and without distractors. Figure 3 shows a schematic co-occurrence matrix of such five word-object pairs, in which each of the first group of words ("Circle" and "Star") has no distractors, and each of the other group of words has two distractors whose referring words are the members of the first group.

Although this statistical regularity in word-to-word relationships looks similar overall to the previous case (compare Figure 2 and 3), this new case is substantially different from the previous cases. The key observation here is that no distractor words have any distractors against themselves. Thus, the first group of words (potential distractors to the other group of words) would be learned via fast mapping, and the other group would be learned also via fast mapping after their distractors are learned before their learning. The learning timing of these two groups are probabilistic, but the first group of words are expected to be learned earlier on average than the other group.

Word	Objects					$\#D$
"Circle"	●					0
"Triangle"	○	▲		☆		2
"Square"	○		■	☆		2
"Star"				★		0
"Diamond"	○			☆	◆	2

Figure 3: A schematic word-object co-occurrence matrix in the case with the two groups of words. Each of the first group of words ("Circle" and "Star") has no distractors, and each of the second group of words ("Triangle", "Square" and "Diamond") has $k = 2$ distractors, whose referring words ("Circle" and "Star") has no distractors.

Efficiency analysis

Specifically, suppose that each word in the group with distractors is learned at the time step t by the probability

$$p_t = (q_t + \overline{q_t}p)\overline{p_{t-1}},$$

where p is the probability to learn this word with distractors at each step, and q_t is the probability to learn it without distractor at step t, or is said the probability for the learning at step t to be *fast mapping.* By setting $\sum_{t=1}^{\infty}(1-p)p^{t-1}t = T_1$ and $q_t = 0$ for any t, this learning time with $k > 0$ distractors is identified with the previous analysis.

Suppose that there are n_0 words without distractors, and $t_0 < t$ samples out of the all $t-1$ samples are drawn from this group of words with equal probability. Then, according to Hidaka (2014), as $n_0 \to \infty$, the probability to learn the m words of this group with the t_0 samples asymptotically approaches to the binomial distribution

$$\sum_{m=0}^{n_0} \binom{n_0}{m} r_t^m \overline{r_t}^{n_0-m}$$

where $r_t := 1 - (1 - 1/n_0)^{t_0}$. If each word in the group with distractors is associated to k distractive words uniformly at random, the fast-mapping probability is

$$q_t = \sum_{m=0}^{n_0} \binom{n_0}{m} r_t^m \overline{r_t}^{n_0-m} \binom{m}{k} / \binom{n_0}{k}.$$

As the hypergeometric distribution[1] approaches the binomial distribution as $n_0 \to \infty$, we obtain

$$\left\| \binom{m}{k} / \binom{n_0}{k} - \binom{m}{k} \left(\frac{k}{n_0}\right)^k \left(1 - \frac{k}{n_0}\right)^{m-k} \right\| \to 0.$$

Using these asymptotic distributions for $n_0 \to \infty$, we obtain the binomial distribution

$$q_t \to \frac{n_0!}{k!(n_0-k)!} \left(r_t \frac{k}{n_0}\right)^k \left(1 - r_t \frac{k}{n_0}\right)^{n_0-k}.$$

With further transform for a sufficiently large n_0, we obtain the fast-mapping probability to be

$$q_t \approx \left(\frac{t_0}{n_0}\right)^k.$$

This expression thus implies that the probability q_t of learning via fast mapping with k distractors approaches 1, if the sample of the words without distractors t_0 is comparable to the number of such words n_0.

Implications

Suppose the number of words without distractors is $n_0 = \gamma n$ with a certain constant $0 < \gamma < 1$, and the number of samples $t_0 = \gamma t$. In this case, as $t_0/n_0 = t/n$, after the point when the number of samples is comparable with the number of words, this learning is sufficiently treated as the fast mapping. Thus, the learning time of a word with k distractors asymptotically approaches the speed of fast-mapping after some constant number of samples for each word. In other words, in the long run, any words would be considered learned in the fast-mapping manner, if any distractor word has no distractors against itself.

This analytic implication is striking in that cross-situational learning on the basis of relative frequency, which itself is as slow as independent learning with a random word-word relationship, can become as efficient as fast-mapping, up to a constant time per word. At the very least, this analysis implies that the nature of the word-to-word relationships is a critical factor in determining the efficiency of relative-frequency based cross-situational learning.

Discussion

In this paper, we studied cross-situational word learning from a theoretical perspective as the formation of a one-to-one word-object map. Our formulation of cross-situational learning is defined as learning on the basis of the relative frequency of objects for each word, which is a more realistic alternative model than eliminative learning, a model analyzed in past studies (Blythe et al., 2010, 2016) that is anyhow a special case of relative frequency learning. Thus, our analysis of relative frequency learning is both more general and more realistic than previously-proposed frameworks. Our analysis shows that its total learning time depends on the minimal difference between the most frequent and the second-most frequent objects among all the words, and that it is quite slow.

Given that relative frequency learning alone is inefficient, we next analyzed the case when the learner applies the lexical constraint that no two referents are associated to a single word. This principle of mutual exclusivity (ME) has been hypothesized to be an important means of reducing ambiguity for children learning language (Markman & Wachtel, 1988; Markman, 1990, 1992), and empirical work has found that both children (Golinkoff et al., 1992; Halberda, 2003; Markman, Wasow, & Hansen, 2003) and adults in cross-situational word learning experiments (Yurovsky & Yu, 2008; Kachergis et al., 2012) show a preference for learning mappings consistent with ME. Using ME, a word can be learned via fast mapping (learned on its first sample), if all the distracting words appearing with it are already learned. However, the effect of ME on the average learning time is quite limited – the same (up to a constant multiplier) as that of independent relative frequency learning, if the distractors for each word are distributed uniformly. In summary, this analysis suggests that the order in which words are learned is related to the statistical nature of the word-to-word relationship–i.e., the structure among the co-occurring distractors.

Therefore, we finally analyzed the case in which a set of words is composed of two word groups: in one group, each word has no distractors, and in the other group each word has k distractors, which are the words without any distractors. Here, it is not just a mixture of two types of words, but the distracting words have no

[1] Gives the probability of k successes in n draws, from a population of size N with exactly K successes. Thus, similar to the binomial distribution, but drawing without replacement.

distractors to themselves, and thus they are likely to be learned earlier than the other group. Thus, in this schematic word structure, the expected learning order is correlated to the number of distractors for the group of words. We hypothesize that, with this statistical regularity, relative frequency learning can be as efficient as learning via fast-mapping, which has been observed in young children (Mervis & Bertrand, 1994; Gershkoff-Stowe & Hahn, 2007). Our analysis suggests that this hypothesis is supported: the learning time is comparable with that of fast mapping learning up to a constant number of samples per word, when a certain ratio of words has no distractors. We expect that this analytic result can be extended to a more general case, such that there are multiple groups with different numbers of distractors up to k and a group of words with k distractors that has no distractors which have k or more distractors against themselves.

In summary, we have analyzed a more general and more realistic class of word learning models, relative frequency learning. Although we showed that learning in this more general framework can be quite slow, we then examined learning under assumptions of mutual exclusivity and word-to-word correlations that might more closely approximate learning situations in the natural language environment. By modifying situations to include realistic variants of these two factors, we showed that learning a full-sized vocabulary could be accomplished on a realistic timescale. Although this work is preliminary, the analytical techniques employed here can be applied to other, yet more realistic cross-situational learning schemes, incorporating better approximations of the language environment, of the problem faced by the learner, and of the biases employed by the learner.

Acknowledgments

This study is supported by the JSPS KAKENHI Grant-in-Aid for Young Scientists JP 16H05860.

References

Akhtar, N., & Montague, L. (1999). Early lexical acquisition: the role of cross–situational learning. *First Language*, *19*, 34–358.

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

Blythe, R. A., Smith, A. D. M., & Smith, K. (2016). Word learning under infinite uncertainty. *Cognition*, *151*, 18–27.

Blythe, R. A., Smith, K., & Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*(4), 620–642.

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

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* Hillsdale, NJ: Erlbaum.

Gershkoff-Stowe, L., & Hahn, E. R. (2007). Fast mapping skills in the developing lexicon. *Journal of Speech, Language, and Hearing Research*, *50*, 682–697.

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

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34.

Hidaka, S. (2014). General type token distribution. *Biometrika*, *101*(4), 999–1002.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin and Review*, *19*(2), 317–324.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57-77.

Markman, E. M. (1992). Constraints on word learning: Speculations about their nature, origins and domain specificity. In M. R. Gunnar & M. P. Maratsos (Eds.), *Modularity and constraints in language and cognition: The minnesota symposium on child psychology* (pp. 59–101). Hillsdale, NJ: Erlbaum.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275.

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name - nameless category (n3c) principle. *Child Development*, *65*, 1646-1662.

Pinker, S. (1984). *Learnability and cognition.* Cambridge, MA: MIT Press.

Quine, W. V. O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. Journal*, *27*, 379–423.

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

Vogt, P. P. (2012). Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognitive Science*, *36*(4), 726–739.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, *4*(1), 32–62.

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. In *Proc. of cogsci 30.* Austin, TX: Cognitive Science Society.

Zipf, G. (1949). *Human behavior and the principle of least effort.* Cambridge, MA: Addison-Wesley.