Cognitive Science 34 (2010) 620–642
Copyright © 2010 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01089.x

Learning Times for Large Lexicons Through Cross-Situational Learning

Richard A. Blythe,[a] Kenny Smith,[b] Andrew D. M. Smith[c]

[a] *SUPA, School of Physics and Astronomy, University of Edinburgh*
[b] *Cognition and Communication Research Centre, Department of Psychology, Northumbria University*
[c] *Language Evolution and Computation Research Unit, Linguistics and English Language, University of Edinburgh*

Received 28 August 2008; received in revised form 29 June 2009; accepted 5 October 2009

Abstract

Cross-situational learning is a mechanism for learning the meaning of words across multiple exposures, despite exposure-by-exposure uncertainty as to a word's true meaning. Doubts have been expressed regarding the plausibility of cross-situational learning as a mechanism for learning human-scale lexicons in reasonable timescales under the levels of referential uncertainty likely to confront real word learners. We demonstrate mathematically that cross-situational learning facilitates the acquisition of large vocabularies despite significant levels of referential uncertainty at each exposure, and we provide estimates of lexicon learning times for several cross-situational learning strategies. This model suggests that cross-situational word learning cannot be ruled out on the basis that it predicts unreasonably long lexicon learning times. More generally, these results indicate that there is no necessary link between the ability to learn individual words rapidly and the capacity to acquire a large lexicon.

Keywords: World learning; Cross-situational learning; Lexicon learning time; Slow mapping; Fast mapping

1. Introduction

Humans excel at learning words—they learn very large vocabularies (around 60,000 words by age 18, or roughly 10 words a day; Bloom, 2000) and can also form an approximate representation of a word's meaning after just a single exposure through *fast mapping* (Carey & Bartlett, 1978; see Horst & Samuelson, 2008; Jaswal & Markman, 2001;

Correspondence should be sent Kenny Smith, Cognition and Communication Research Centre, Department of Psychology, Northumbria University, Northumberland Building, Northumberland Road, Newcastle upon Tyne NE1 8ST, UK. E-mail: kenny.smith@northumbria.ac.uk

Wilkinson & Mazzitelli, 2003; Woodward & Markman, 1998 for reviews). A causal relationship between these phenomena is widely assumed, and there are suggestive correlations between the onset of the ability to fast map and the time at which vocabulary begins to rapidly expand (summarized in Wilkinson & Mazzitelli, 2003, pp. 48–49, but see McMurray, 2007 for an alternative explanation of the vocabulary explosion).

However, the process of fast mapping a new word represents the start, not the end, of word learning: The approximate word meanings established by fast mapping need to be fleshed out through a process dubbed *slow mapping* by Carey (1978), involving identifying a word's extension, elaborating its meaning, and placing it within the broader semantic network (see McGregor, 2004, for a useful summary). Indeed, Carey's influential account suggests that the initial fast mapping event establishes little more than a placeholder in the lexicon that forms the basis for this subsequent slow mapping process. Recent work further suggests that these fast-mapped lexical entries may be very fragile indeed and prone to being forgotten unless bolstered by environmental cues that support the immature lexical entry (Horst & Samuelson, 2008). The implications of the more gradual nature of slow mapping for the learning of large lexicons are unclear: While it seems obvious that rapidly adding words to the lexicon via fast mapping will facilitate learning large lexicons, the same logic suggests that the slow mapping process will potentially limit the eventual size of the lexicon attained.

Why is slow mapping necessary? In other words, why are the representations of word meaning established by fast mapping incomplete approximations? One-shot word learning is problematic because it requires a word learner to accurately infer the meaning of a new word the first time he or she hears it. This is not straightforward: As noted by Quine (1960), there are in principle infinitely many possible meanings that would be consistent with a particular utterance (or sequence of utterances) of a word. He imagined an anthropologist interacting with a native speaker of an unfamiliar language. As a rabbit runs by, the speaker exclaims ''gavagai,'' and the anthropologist notes that ''gavagai'' means *rabbit*. Quine showed, however, that the anthropologist cannot be sure that ''gavagai'' means *rabbit*; in fact, it could have an infinite number of possible meanings, such as *undetached rabbit parts*, *dinner*, or even (perhaps a superstition of the speaker) *it will rain*.

This infinite range of possible meanings must be reduced to a more manageable size in order for word learning (via slow or fast mapping) to be possible. Various sociopragmatic, representational, interpretational, and syntactic *heuristics* have been proposed to explain how this might be achieved: Children use behavioral cues to identify the attentional focus of a speaker in order to infer word meaning (Baldwin, 1991; Tomasello & Farrar, 1986); children assume that words refer to whole objects, rather than parts or properties of those objects (Landau, Smith, & Jones, 1988; Macnamara, 1972); knowledge of the meaning of other words is used to infer the meaning of a new word, for example, by assuming that words have mutually exclusive meanings (Markman & Wachtel, 1988); argument structure and syntactic context facilitate word learning, particularly for ''hard words'' such as verbs denoting abstract relationships (Gillette, Gleitman, Gleitman, & Lederer, 1999; Gleitman Cassidy, Nappa, Papafragou, & Trueswell, 2005). In order for a word's meaning to be learned in a single exposure, these various *word learning heuristics* would have to act in concert to uniquely and reliably identify the meaning of the word being learned. This is a demanding task, requiring strong heuristics. Could a large lexicon still be learned if the learner's heuristics were somewhat weaker, and sometimes (or even routinely) failed to eliminate all uncertainty as to a word's meaning?

Cross-situational learning is a mechanism for word learning in the face of this kind of *referential uncertainty*. The idea behind cross-situational learning (as discussed in e.g., Pinker, 1989, 1994) is that the context of use (in conjunction with the learner's word learning heuristics) provides a number of candidate meanings for a word, each of which is in principle equally plausible. If the same word is produced in a different situation, a different set of candidate meanings may be suggested. The learner can make use of this cross-situational information—the true meaning of the word will lie at the intersection of the two sets of candidate meanings—and repeated exposure therefore enables the learner to reduce his or her uncertainty as to the word's true meaning. As such, cross-situational learning falls within the much larger set of processes involved in slow mapping: It is one mechanism by which a learner can refine his or her understanding of a word's meaning over time.

Experimental studies involving the acquisition of small numbers of words from sequences of artificial or naturalistic exposures suggest that humans (both adults and infants) are capable of cross-situational learning (Akhtar & Montague, 1999; Gillette et al., 1999; Smith & Yu, 2008; Yu & Smith, 2007, but see Smith, Smith, & Blythe, 2009 for a critique of the methodology employed by Yu & Smith, 2007). Formal models (reviewed in Section 2) also suggest that cross-situational learning can be used to accurately infer the meanings of words from corpora. Existing formal models typically focus on showing that a cross-situational learner can accurately learn the meaning of a relatively small set of words from a small (but realistic) corpus of language use. This is a worthwhile and important enterprise. However, these models do not at present show that cross-situational learning can scale up to the learning of human-sized vocabularies. In Section 3, we show, via a mathematical model, that such scaling is in principle possible—there is no necessary link between rapidly learning the meaning of individual words and eventual acquisition of large vocabularies, and cross-situational learning potentially facilitates the rapid acquisition of large vocabularies despite massive levels of referential uncertainty. While our formal model deals with a much more stereotyped and simplified word learning scenario, this result suggests that it is worth pursuing these more realistic formal models on increasingly complex corpora. The results of this model also have more general implications for the relationship between speed of learning individual words and eventual vocabulary size. As we discuss below, our general technique could be used to derive an estimate of overall lexicon learning times for any theory of word learning that provides an estimate of learning times for single words.

2. Existing formal treatments of cross-situational learning

Siskind (1996) presents an early and influential operationalization of cross-situational learning, providing an algorithm capable of correctly extracting word meanings from a synthesized corpus of utterances paired with (intended and spurious) meanings, despite referential uncertainty, homonymy, and noise. Siskind's cross-situational learner proceeds via the eliminative process outlined above, attempting to identify a word's meaning by winnowing down a set of candidate word meanings across exposures. Siskind also shows that cross-situational learning procedures can be specified in such a way as to allow a learner to retreat from errors introduced by environmental noise or homonymy. For example, a common criticism of the eliminative cross-situational learning algorithm (see e.g., Gleitman, 1990) is that it breaks down in situations where the intended referent for a word is not present in the situation in which the word is uttered—in such a scenario, a strict eliminative learner will rule out the word's true meaning due to this noisy data point. Similarly, two homonymous words will share a null intersection of meaning, as there will be no common meaning consistently present across multiple uses of those homonyms. Siskind's learning algorithm is capable of identifying and correcting these sorts of errors (by associating confidence scores with word-meaning associations, and allowing back-tracking and splitting of lexical entries based on those confidence scores).

In addition to his basic finding that working cross-situational learning algorithms can be provided, Siskind also provides a limited sensitivity analysis in an attempt to identify how his algorithm copes with increasing task difficulty along several dimensions. Siskind reports, based on a small number of simulation runs, that lexicon learning time:

1. increases approximately linearly with lexicon size;
2. increases as noise or degree of homonymy in the target lexicon increases;
3. is invariant with respect to the number of conceptual primitives used to construct utterance meanings; and
4. is invariant with respect to degree of referential uncertainty at each exposure.

His third and fourth findings are particularly surprising in the context of the theoretical debate on cross-situational learning. For example, it is often assumed that increases in degree of representational complexity and referential uncertainty will lead to some sort of explosion of complexity which will necessarily stymie the process: ''the trouble is that an observer who notices *everything* can learn *nothing*, for there is no end of categories known and constructable to describe a situation'' (Gleitman, 1990, p. 12); ''The very richness of perception guarantees multiple interpretative possibilities at many levels of abstraction for single scenes; but the problem for word learning is to select from among these options the single interpretation that is to map on to a particular lexical item'' (Gleitman, 1990, p. 13). This point is generally immediately conceded even by proponents of cross-situational learning (e.g., by Pinker, 1994, see p. 392). Yet Siskind's finding seems to suggest that an explosion of complexity is not inevitable—neither a proliferation of conceptual primitives, nor an increase in the level of referential uncertainty per exposure produces, at least for his algorithm, *any* decrease in performance. It seems important to explore whether Siskind's finding is generally true, or whether it is perhaps an artifact of his model or a consequence of the fairly limited nature of his sensitivity analysis.

More recent formal models of cross-situational learning have adopted more probabilistic notions of the meaning-form mapping in the lexicon and have ratcheted up the level of realism of the data that the cross-situational learner is exposed to. Yu, Ballard, and Aslin (2005) describe an impressive system that takes video of visual scenes paired with natural-language audio descriptions of those scenes as input and develops a lexicon of associations between visual objects (parsed out from the visual scene) and spoken words (segmented from the speech stream). At the heart of this model lies a cross-situational learning mechanism that stores a lexicon as a set of probabilistic associations between words and objects and calculates the lexicon that best accounts for the cross-situational usage data. Despite a highly complex set of input stimuli, this system correctly identifies the meaning of approximately 70% of the word-object pairings present in its input. Similarly, working in a Bayesian framework, Frank, Goodman, and Tenenbaum (2009) present a model that proceeds from real child-directed speech data paired with a manually produced description of the contents of the associated scenes to successfully learn small lexicons.

These models show great promise for the development of systems capable of cross-situational word learning from real-world data. However, they are at present only applied to small (though relatively complex and realistic) corpora, involving a limited number of possible referents and a limited lexicon. The development of this sort of system, as an existence proof for the viability of cross-situational word learning in environments of high complexity, strikes us as an extremely important one. However, it presupposes that there is no fundamental cutoff point at which an increase in lexicon size, semantic or environmental complexity, or referential uncertainty will render cross-situational learning impossible. Siskind's sensitivity analysis offers some positive indications that this faith is justified, but given the complexity of his algorithm, his analysis is necessarily rather sparse. It is therefore desirable to place cross-situational learning on a more solid theoretical footing: As well as showing that it can be made to work for increasingly complex corpora, can we be confident that there is no lurking performance ceiling that will limit cross-situational learning to (relatively) toy worlds? The mathematical analysis that follows is an attempt to address such a question.

3. Learning time for a simple model lexicon

3.1. Rationale

Our primary aim in this paper is to understand how referential uncertainty affects the time taken to learn a lexicon of human proportions. For this purpose, we introduce an idealized mathematical model that allows us to calculate and compare the time required to acquire a large lexicon through cross-situational learning under a variety of degrees of referential uncertainty. We stress that this model is not intended to provide a cognitively plausible account of cross-situational word learning: The models reviewed above (particularly Yu et al., 2005 and Frank et al., 2009) are much more sophisticated in this regard. Rather, our aim is to provide an initial evaluation of whether cross-situational learning can in principle scale up to the learning of large lexicons, and whether there is any inherent cutoff point of referential uncertainty or lexicon size at which cross-situational learning becomes impossible. This necessitates formulating a much simpler treatment of cross-situational learning, at least at first—ideally this can then be elaborated to provide a similar evaluation of the cross-situational algorithms provided by Siskind (1996), Yu et al. (2005), and Frank et al. (2009). We begin by defining our model and the assumptions that go into it, and then we return in the discussion to the limitations of these assumptions and the likely consequences of relaxing them.

3.2. Definition of the model

The model lexicon comprises *W* words, each of which has a unique meaning. The learning agent experiences a sequence of *learning episodes*. In each of these episodes, a single target word is presented (e.g., spoken) to the learning agent. Whenever the target word is presented, its associated *target meaning* is assumed always to be present (i.e., inferrable from the context in which the word is uttered). Alongside the target meaning (i.e., also inferrable) are a number of other *incidental meanings*. Although, as discussed above, there may be infinitely many of these incidental meanings, we assume that the learning agent is equipped with some algorithm (i.e, the heuristics discussed earlier: attentional focus of speaker, whole object bias, etc.) to reduce the number of candidate meanings present in a given episode to a finite (and possibly small) number.[1] There are two key parameters that enter into the model here: *M* is the number of incidental meanings that *might* be inferred alongside the true target meaning; *C* is the number of incidental meanings that *are* inferred in a given episode. This latter set comprises those meanings that were not eliminated by the learner's heuristics (see Fig. 1).

By definition, $0 \leq C \leq M$. Application of powerful word-learning heuristics will eliminate incidental meanings and lead to small *C*, whereas weaker heuristics will leave greater uncertainty and larger *C* (Golinkoff, Mervis, & Hirsh-Pasek, 1994). The ratio *C*/*M* quantifies the strength of these heuristics, the degree of uncertainty, and hence the difficulty of the problem the learner has to solve. If this ratio is large, incidental meanings may consistently

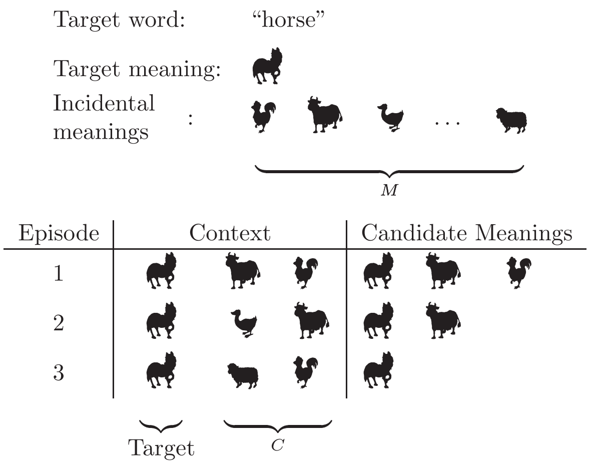

Fig. 1. Cross-situational learning of the meaning of *horse*, with $C = 2$. Given the particular sequence of exposures illustrated here, the word is learned on the third episode.

appear alongside the target meaning and thus be plausible (though incorrect) candidates for the word's meaning, thus delaying word learning.

In order to make progress in analyzing the performance of specific learning strategies under different degrees of uncertainty, we make a number of simplifying assumptions. First, we take the values of C and M to be the same for each target and episode. In any given episode, the C incidental meanings are drawn uniformly at random, and without replacement, from the full set of M meanings associated with the target word. This sampling is assumed to be independent in each episode (i.e., a given incidental meaning has the same probability of appearing whenever an associated target is presented). The target word itself is also selected at random from all possible words in the lexicon, but not necessarily uniformly. To this end we introduce the probability ϕ_i that word i is presented in a given episode. Again, each presentation is a statistically independent event: Bursts and lulls in the temporal distribution of words that have been reported elsewhere (Altmann, Pierrehumbert, & Motter, 2009) are not included in this first model.

Note that we do *not* assume any relationship between the sets of M incidental meanings associated with different target words. There may be complete overlap between some sets of incidental meanings (for example, when the targets are very similar) or no overlap at all. The results we obtain below are independent of such considerations. Moreover, our discussion of meanings as unstructured, atomic entities is purely for ease of exposition. Within this model, meanings could equally be interpreted as existing in a hierarchically and similarity-structured space. This structure would be reflected in the set of incidental meanings associated with each target meaning and the distribution from which those incidental meanings are drawn, such that similar meanings tend to occur in one another's incidental meaning sets and more similar and more general meanings tend to be selected as distractors more frequently.

The final assumption we make is that words are learned independently. That is, once the meaning of one word has been established, that knowledge is not then used by the learner to make inferences about possible meanings of other words: for example, we do not assume that learners apply a mutual exclusivity constraint (Markman & Wachtel, 1988). This assumption of independence implies that the learning time for a lexicon can be determined from the learning time for a single word (see below).[2] While we return to this issue in the discussion, for the moment we merely reemphasize that this model is intended as a simple sketch, rather than a realistic, exhaustive treatment.

3.3. Learning times for the model lexicon

We now calculate the time taken for a learner to acquire the lexicon of W words under the conditions described above using three, progressively weaker, word-learning strategies.

3.3.1. Lexicon learning times for a one-shot word learner

Let us first take the case of a learner who can identify the target meaning for a word on his or her first encounter with that word—the most powerful form of fast mapping possible. In order to achieve this, all incidental meanings must be eliminated by the learner's heuristics, and $C = 0$. In order for this learner to learn the entire lexicon, each of the *W* words must have been presented at least once. In principle this could be achieved in $t = W$ episodes, but this will in general not happen: Given each word could be repeated arbitrarily many times, there is some probability that at any finite time *t*, at least one word in the lexicon has never been presented to the learner and therefore has not been learned. Our definition of a lexicon learning time must therefore be probabilistic. We thus introduce $P_W(t)$, the probability that all *W* words have been learned by time *t*. We deem the lexicon to be learned when this probability is sufficiently close to unity, that is, when $P_W(t) = 1 - \epsilon$ with ϵ a small parameter. The time at which this occurs we denote as t^*. For example, $\epsilon = 0.01$ means that the lexicon has been learned with 99% probability; or equivalently, that if 100 agents are learning the lexicon in parallel, but from different sequences of exposures, all but one of them are expected to have learned all *W* words by time t^*.

A quick way to estimate the learning time t^* when each word is equally likely to be presented in each episode is as follows. Let $u(t)$ be the expected number of words that remain to be learned at episode *t*. This number decreases at a rate equal to the probability that a previously unheard word is exposed in the next episode. Since all words appear in each episode with equal probability, this probability is $u(t)/W$. Hence,

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = -\frac{u(t)}{W}. \tag{1}$$

This differential equation has the solution

$$u(t) = W\mathrm{e}^{-t/W}, \tag{2}$$

given that at $t = 0$, all *W* words remain to be learned. If *W* is large, the learning time t^* will also be large (since we know $t^* \geq W$). At very large times, the most likely number of words that remain to be learned is either zero or one; hence, at these times the expected number of unlearned words equals the probability the lexicon has not been learned, that is,

$$\epsilon = 1 - P_W(t^*) \approx u(t^*) = W\mathrm{e}^{-t^*/W}. \tag{3}$$

Rearranging this expression gives an estimate for t^* for a one-shot, fast-mapping learner as

$$t^*_{\mathrm{FM}}(\epsilon) \approx W \ln\left(\frac{W}{\epsilon}\right). \tag{4}$$

That is, the typical number of episodes required until the lexicon is learned is far greater than the size of the lexicon, purely as a consequence of having to wait for unseen words to appear. For example, in the case $\epsilon = 0.01$, a lexicon of the size typical for a human adult, $W \approx 60{,}000$, and a uniform word distribution, requires about 940,000 exposures to be learned by a learner capable of learning each word after just one exposure. While the required number of exposures is large relative to the size of the lexicon, it is extremely small relative to the number of words children are likely to encounter in a day. For instance, this amounts to a modest 142 learning episodes (i.e., encounters with words) per day for 18 years, well below the 600–2,100 words *per hour* likely to be spoken by parents to children (Hart & Risley, 2003). In other words, one-shot learning is far more powerful than required to learn a lexicon in a practicable timescale, suggesting that lexicon learning times for less powerful learning strategies should be quantified.

3.3.2. Developing a general formulation for lexicon learning time

Similar expressions to (4) are obtained for more general word distributions, and for values of $C > 0$ (i.e., when the target meaning cannot be identified on a word's first exposure). The reason for this is that in each case, the probability that the lexicon has not been learned decays exponentially to zero at large times; rearranging this exponential then results in an expression of the form (4), albeit with different constants appearing that depend on the learning strategy, degree of referential uncertainty, and word distribution.

More precisely, we show in Appendix A how to relate $P_W(t)$, the probability that *all* W words have been learned after t exposures, to $P_1(t)$, the corresponding quantity for a *single* word. It turns out that all the learning strategies we consider below can be analyzed through a generic expression for the single-word learning function

$$P_1(t) = \begin{cases} 0 & \text{if } t = 0 \\ 1 - a(1-q)^t + r(t) & \text{for } t > 0 \end{cases} \tag{5}$$

that contains two parameters a and q that depend on the strategy and will be related to M and C below for specific strategies. The general features of this function are as follows. (i) The learner always learns the correct meaning of a word given enough exposures: as $t \to \infty$, $P_1(t) \to 1$. (ii) The parameter q quantifies the late-time behaviour of the learning algorithm: It is the rate at which the word is learned after many exposures *given* that it has not yet been learned (e.g., due to the presence of many confounding meanings). (iii) Meanwhile, the early-time behaviour of the algorithm is rolled into the single parameter a. If a is small, the word is likely to have been learned in the first few episodes; by contrast if it is large, it is unlikely to have been learned quickly. Note that the early-time shape of the single-word learning function may be very complicated: Its details turn out to be irrelevant to the overall learning time for a large lexicon, as long as a technical assumption on the remainder term $r(t)$ is satisfied, namely that $\lim_{t\to\infty} r(t)(1-q')^{-t} \to 0$ for some $q' > q$. This assumption is valid for all the cases we consider here.

The result derived in Appendix A is that, for sufficiently large t,

$$P_W(t) \sim \prod_{i=1}^{W} [1 - a\mathrm{e}^{-\phi_i q t}], \tag{6}$$

where we recall that ϕ_i is the exposure frequency of word i. For the simple case of a uniform distribution, $\phi_i = 1/W$, we find

$$P_W(t) \sim [1 - a\mathrm{e}^{-qt/W}]^W. \tag{7}$$

Setting this equal to $1 - \epsilon$ and inverting, we obtain an estimate for the lexicon learning time:

$$[1 - a\mathrm{e}^{-qt^*/W}]^W = 1 - \epsilon \tag{8}$$

$$\Longrightarrow 1 - a\mathrm{e}^{-qt^*/W} = (1 - \epsilon)^{1/W} \tag{9}$$

$$\Longrightarrow \mathrm{e}^{-qt^*/W} = \frac{a}{1 - (1 - \epsilon)^{1/W}} \tag{10}$$

so then, after taking the logarithm on both sides, we find

$$t^* \approx \frac{W}{q} \ln\left(\frac{a}{1 - (1 - \epsilon)^{1/W}}\right) \approx \frac{W}{q} \ln\left(\frac{aW}{\epsilon}\right). \tag{11}$$

The second approximate equality holds if ϵ is small or W is large (both of which correspond to regimes of interest).

Let us return to the previous example of fast mapping. Here the appropriate choice for the parameters a and q are $a = q = 1$. Then, we have from (5) that $P_1(0) = 0$ and $P_1(t) = 1$ for $t > 0$ if $r(t) = 0$. That is, (5) gives the single word learning probability function exactly, since in this case we assume that the word is learned immediately on its first exposure. Substituting these values into (11) recovers the expression (4) previously obtained by other means.

3.3.3. Lexicon learning times for a proficient cross-situational learner

We are now equipped with the tools needed to examine the performance of a *pure* cross-situational learner, that is, an agent who admits only those meanings that have appeared in *all* previous episodes involving the target word as possible candidates for its true meaning. Over time, the size of the set of candidate (but incorrect) meanings decreases to zero: As soon as an incidental meaning fails to appear, it can be excluded as a candidate meaning. The rate of this decrease is controlled by the parameters C and M: If C is small relative to M, meanings are excluded rapidly. We showed in a previous work (Smith, Smith, Blythe, & Vogt, 2006) that the probability that this set comprises k meanings after t exposures of the target is

$$R_k(t) = \binom{C}{k} \sum_{r=k}^{W} (-1)^{k-r} \binom{C-r}{k-r} p_r^{t-1} \tag{12}$$

where

$$p_r = \frac{\binom{M-r}{C-r}}{\binom{M}{C}}. \tag{13}$$

If forced to guess the correct meaning of the word, the only rational behaviour for the agent is to choose at random from the set of $k + 1$ meanings that have always appeared alongside the word. If we use the probability of a correct guess after episode t to define the probability of having learned the word, we find that

$$P_1(t) = \sum_{k=0}^{C} \frac{1}{k+1} R_k(t) = \sum_{r=0}^{C} \frac{(-1)^r}{r+1} \binom{C}{r} p_r^{t-1}, \tag{14}$$

where the second equality emerges after some manipulation. We remark that if agents employ the ‘‘guess-and-test’’ strategy that we have observed in word-learning experiments (K. Smith, A.D.M. Smith, & R.A. Blythe, unpublished data), whereby they form a hypothesis for the target meaning by choosing from the $k + 1$ candidate meanings *and maintain* that hypothesis until such time as that meaning is absent, $P_1(t)$ corresponds exactly to the probability that the agent holds the correct hypothesis after t episodes.[3]

Comparing this expression with (5), we identify $a = M/2$ and $q = 1 - (C/M)$, and hence that the time needed to learn a large lexicon under cross-situational learning (XSL) is

$$t^*_{\mathrm{XSL}}(\epsilon) \approx W \frac{1}{1 - \frac{C}{M}} \ln\left(\frac{MW}{2\epsilon}\right), \tag{15}$$

when target words are selected according to a uniform distribution.

We can see from Fig. 2 that, for example, the cross-situational learning time for the case C 17 and $M = 100$ is only 50% longer than that of a fast mapping learner, and at 214 learning episodes per day still represents only a tiny fraction of the words heard every day by the aver-

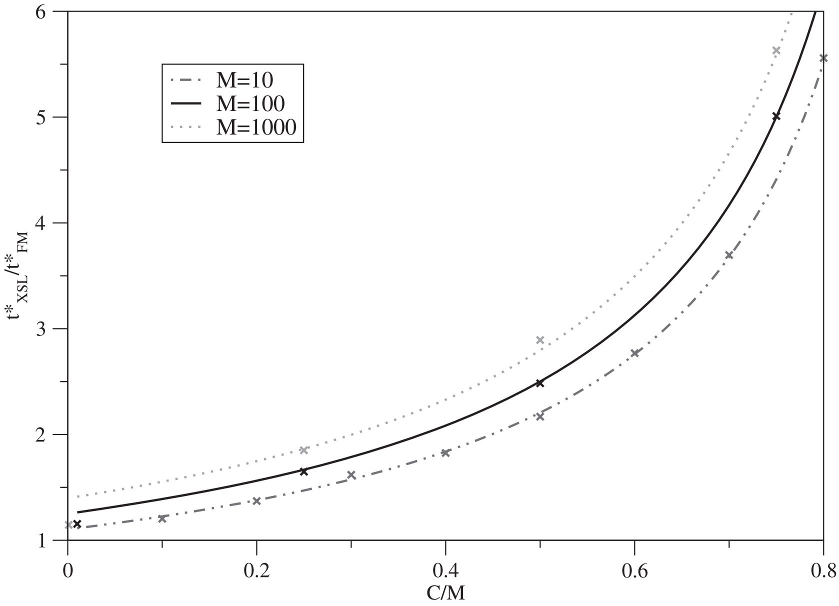

Fig. 2. Cross-situational learning times as a function of *C/M*, for a uniform target word distribution and various values of *M*, as a proportion of time taken by a fast-mapping learner ($t^*_{\mathrm{XSL}}/t^*_{\mathrm{FM}}$), with $\epsilon = 0.01$; the corresponding curves for a Zipfian target word distribution are indistinguishable. Points show the time required for a proportion $1 - \epsilon$ of learners to learn the whole lexicon for a sample of 2,000 Monte Carlo simulations of the learning process (see Appendix B).

age child, according to the figures provided by Hart and Risley (2003). While comparison to this real-world estimate is of limited utility unless we know the real values of parameters C and M (an issue we return to in the discussion), the important point is that the increase in learning times associated with cross-situational learning is (under a large portion of the parameter space) rather modest relative to the learning times provided by one-shot learning.

It is also useful to relate our findings to Siskind's (1996) more limited analysis. For example, Siskind's conclusion that the lexicon learning time increases approximately linearly with lexicon size is confirmed by our calculation: The time grows generically as $W \ln W$, which empirically is almost indistinguishable from a linear growth. While the conclusion that degree of referential uncertainty (C/M in our model) and conceptual complexity (M in our model) have *no* impact on lexicon learning times is not supported, their impact is certainly small, particularly at the low levels that Siskind explored in his sensitivity analysis. Only when C/M approaches 1 do lexicon learning times for efficient cross-situational learners explode.

It is worth highlighting the limitations of this approximate formula. A comparison of our formula with the results from Monte Carlo simulations of the learning strategies (which are exact, up to sampling errors) reveals that the learning time is overestimated at small C/M. This can be seen from Fig. 2, where the crosses obtained by simulation lie just below the curves as $C/M \to 0$. The reason for this discrepancy is that the correction term $r(t)$ in Eq. 5 can no longer be neglected (see Appendix A). We note in particular that the result is invalid for the case $C = 0$, where we have shown that Eq. 4 is the correct expression.

3.3.4 Lexicon learning times for a limited cross-situational learner

In this previous calculation, we have assumed that learners can make maximum use of cross-situational information, that is, they can maintain an accurate set of candidate meanings for each word (those meanings that consistently occur with the word), as well as their preferred candidate hypothesis from this set. We can also identify a strategy that makes *minimal* use of cross-situational information, that is, where only one candidate hypothesis for the word's meaning is taken forward from one exposure to the next. We assume, as with the ''guess-and-test'' strategy previously described, that this hypothesis is changed when the meaning in question fails to appear with the target word, at which point a new candidate meaning is selected at random from the set of meanings co-occurring with the target word, and without reference to any earlier exposures to that word. This new hypothesis is subsequently maintained until such times as it too is proven to be incorrect, and so on.

Let $Q_1(t) = 1 - P_1(t)$ be the probability that the agent holds an incorrect hypothesis after t exposures. Under the conditions we have described, it is impossible to switch away from the correct hypothesis. On the other hand, a switch from an incorrect hypothesis to the correct one is possible, and indeed, the probability of this event is the same in each episode. First, a change of hypothesis occurs if the previous hypothesis failed to appear; this happens with probability $\frac{M-C}{M}$. Secondly, the new, randomly chosen, hypothesis is correct with probability $\frac{1}{C+1}$. The total probability of identifying the correct hypothesis on an episode given that the current hypothesis is incorrect is thus the product of these two probabilities, $\frac{M-C}{M(C+1)}$; the probability that the hypothesis is still false at time $t + 1$ is thus

$$Q_1(t+1) = \left[1 - \frac{M-C}{M}\frac{1}{C+1}\right]Q_1(t) = \frac{M+1}{M}\frac{C}{C+1}Q_1(t), \tag{16}$$

where the second equality follows after rearrangement. This reveals that the probability of holding a false hypothesis decreases by the same factor in each episode. Hence,

$$Q_1(t) = \left[\frac{M+1}{M}\frac{C}{C+1}\right]^{t-1} Q_1(1). \tag{17}$$

The probability of being incorrect after the first exposure, $Q_1(1)$ is $\frac{C}{C+1}$ (since C of the $C + 1$ choices are incorrect). Therefore,

$$Q_1(t) = \left[\frac{M+1}{M}\frac{C}{C+1}\right]^{t-1} \frac{C}{C+1} = \frac{M}{M+1}\left[\frac{M+1}{M}\frac{C}{C+1}\right]^{t}. \tag{18}$$

By using the fact that $P_1(t) = 1 - Q_1(t)$ we find an expression that is once again of the standard form (5):

$$P_1(t) = 1 - \frac{M}{M+1}\left[\frac{M+1}{M}\frac{C}{C+1}\right]^{t}. \tag{19}$$

The parameters a and q are $a{=}M/(M{+}1)$ and $q{=}(1{-}C/M)/(C{+}1)$. From (11), we find that this minimally cross-situational strategy (''min'') leads to the learning time

$$t^*_{\min} \sim \frac{W(C+1)}{1-\frac{C}{M}} \ln\left(\frac{MW}{(M+1)\epsilon}\right) \tag{20}$$

that is approximately $C + 1$ times longer than that for pure cross-situational learning. This highlights the extent to which good use of cross-situational information can accelerate lexicon learning.

3.3.5. Lexicon learning times for a frequentist cross-situational learner

Any strategy that is more effective than the minimal strategy presented in the preceding section (which we dub Minimal XSL), but less effective than the fully eliminative cross-situational behaviour described in Section 3.3.3. (Pure XSL), will necessarily have $P_1(t)$ greater than that for Minimal XSL, but less than that for Pure XSL, for any t. Translated into learning times, this implies that

$$t^*_{\mathrm{XSL}} \le t^* \le t^*_{\min} \tag{21}$$

for any intermediate strategy that is consistent with the assumptions of the previous section. In particular, this includes a refinement of the minimal strategy in which agents select a hypothesis meaning not uniformly at random from all meanings present, but with a probability proportional to the number of times it has appeared alongside all exposures of the target word to date. We have found this probabilistic strategy (Approximate XSL) to provide a good fit to experimental data (K. Smith, A.D.M.

Smith, & R.A. Blythe, unpublished data). In the absence of an exact formula for $P_1(t)$ for this strategy, we have estimated its associated lexicon learning times by means of Monte Carlo simulations. The results, along with the bounds imposed by the Pure and Minimal XSL strategies, are shown in Fig. 3. As expected, Approximate XSL falls between Pure and Minimal XSL. As we can see, even the weaker forms of cross-situational learning still allow the acquisition of large vocabularies in practicable timescales despite considerable uncertainty (perhaps even up to $C/M \approx 0.7$) at each exposure.

3.3.6. Additional observations

We conclude this section with two further observations. First, Eq. 7 gives an expression for the lexicon learning probability for nonuniform word distributions. In particular, we may consider the Zipfian distribution (Zipf, 1949), in which the frequency of the *n*th most common target word is proportional to $1/n$ (note, however, that the C coincidental meanings are still assumed to be uniformly sampled from the M possibilities, a point we return to in the discussion). If $t \gg 1/(\phi_{\min} q)$ we may legitimately write that

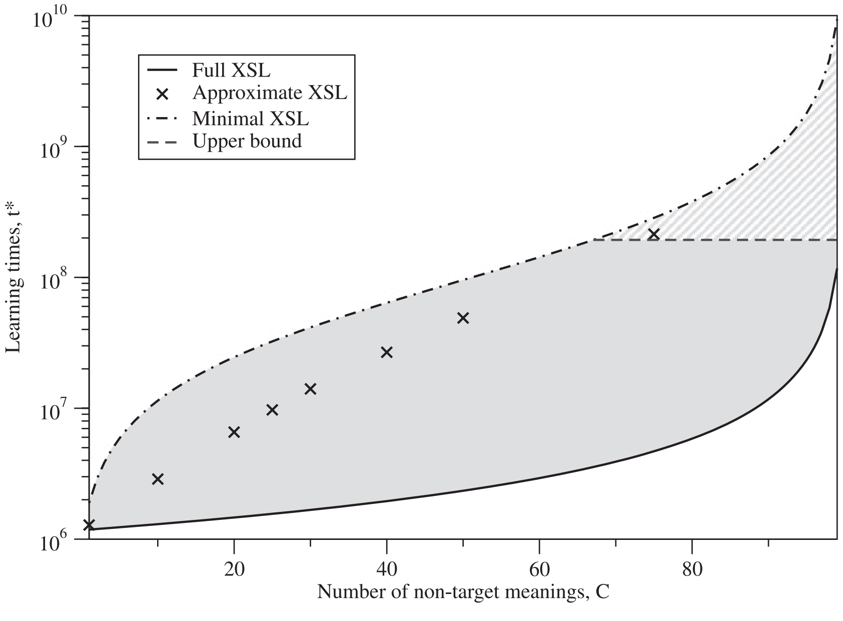

Fig. 3. Learning time as a function of C, for pure cross-situational learners (solid line) and Minimal XSL (chain line), for $\epsilon = 0.01$, $M = 100$, $W = 60{,}000$. These two strategies give lower and upper bounds, respectively, on cross-situational learning time—all XSL strategies will fall within the shaded region (e.g., Approximate XSL, given by points). The blue horizontal line gives an extrapolation from the number of exposures suggested by Hart and Risley (2003): 2,100 words per hour, 14 h of exposure per day for 18 years—this probably represents an upper bound on the true figure. The hatched region indicates values of C that would render a lexicon of 60,000 words unlearnable via any of our cross-situational technique in this time limit.

$$\ln P_W(t) \sim \sum_{i=1}^{W} \ln\left[1 - a\mathrm{e}^{-\phi iqt}\right] \approx -a \int_1^W \mathrm{d}x \exp\left(-\frac{qt}{\mu x}\right) \tag{22}$$

where $\mu = \sum_{i=1}^{W} 1/i$. One can rewrite this expression in terms of exponential integral functions whose asymptotic behavior for large argument is known (Abramowitz & Stegun, 1965). Keeping the largest terms in the asymptotic expansions finally leads to

$$\ln P_W(t) \approx \frac{-a\mu W^2 \mathrm{e}^{-qt/\mu W}}{q_1 t}. \tag{23}$$

Setting $P_W(t)$ equal to $1 - \epsilon$ and inverting, as before, leads to the formula

$$t^* \approx \frac{W\mu}{q} \mathscr{W}_0\left(-\frac{aW}{\ln(1-\epsilon)}\right) \tag{24}$$

in which $\mathscr{W}_0$ is the principal branch of Lambert's W function (Corless, Gonnet, Hare, Jeffrey, & Knuth, 1996).

The main thing to be aware of is that this function behaves for large argument as a logarithm. Thus, for small ϵ, the only real difference between this expression and (11) is the factor μ. For the lexicon size $W = 60{,}000$, we find that $\mu = 11.579\ldots$, and hence for all strategies whose single-learning function can be expressed in the form (5), we expect the learning time for a Zipf-distributed lexicon is increased by a factor of μ over that for a uniformly distributed lexicon. Note in particular that this increase in learning time is predicted to be independent of C and M. This prediction is confirmed by the Monte Carlo simulation data, shown in Fig. 4. Furthermore, this implies that the performance of cross-situational learning relative to one-shot learning is therefore the same for both uniform and Zipfian distributions of target words. We remark that the absolute increase in learning times for the Zipfian distribution is very modest, given that the rarest word is uttered 60,000 times less frequently than the most common.

Our second observation is that this general approach to deriving lexicon learning times from an account of individual word learning is not restricted to cross-situational learning. In principle, we can provide equivalent expressions for any theory that specifies the speed of individual word learning. More generally, all theories of word learning contain an implicit prediction regarding the number of exposures required to learn a large lexicon, which can be made explicit by instantiating that theory in a model and calculating lexicon learning times under that model. This potentially provides an additional means of evaluating such theories: Our calculations suggest that cross-situational learning cannot be rejected on the basis that it predicts unreasonably long learning times for large lexicons.

4. Discussion

In the previous section we have shown that, under rather idealized conditions, cross-situational information allows a learner to achieve learning rates comparable to those

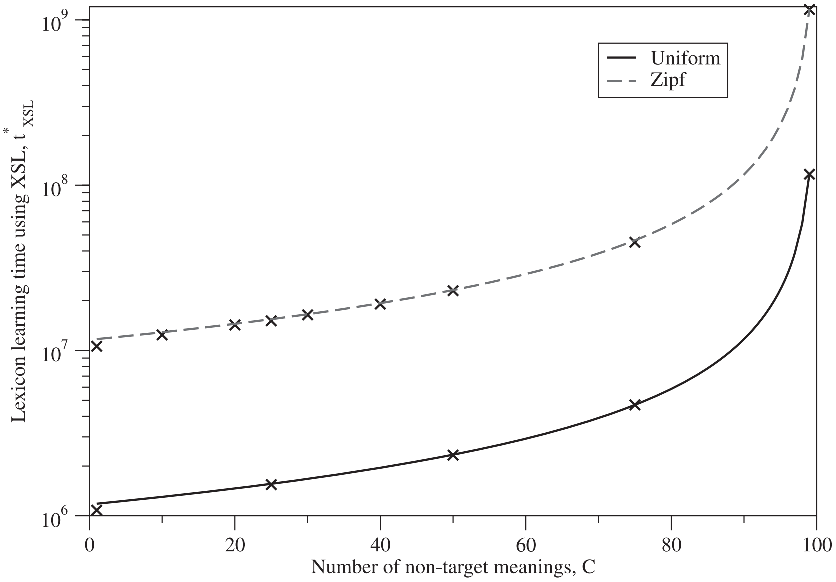

Fig. 4. Cross-situational learning time as a function of *C*, for a lexicon of 60,000 words ($W = 60{,}000$), in which there are $M = 100$ incidental meanings for each word, and $\epsilon = 0.01$. The solid line is for a uniform target word distribution, and the dashed line is for a Zipfian distribution. Points show the time required for a proportion $1 - \epsilon$ of learners to learn the whole lexicon for a sample of 2,000 Monte Carlo simulations of the learning process (see Appendix B).

obtained in the absence of referential uncertainty (when it becomes possible to learn each word after a single encounter), even in the presence of a large degree of uncertainty at every exposure to each word. Unsurprisingly, lexicon learning is fastest when word learning heuristics are strong enough to eliminate all uncertainty as to word meaning. However, cross-situational learning is still possible when these heuristics are weaker and admit a far greater degree of uncertainty as to word meaning. In other words, there is *no* necessary link between the ability to rapidly learn individual words and the ability to acquire large vocabularies: Vocabularies on the human scale can be acquired relatively rapidly by a proficient cross-situational learner. We note further that, given that learning words rapidly requires the elimination of all uncertainty as to word meaning, which is likely to require sophisticated and cognitively demanding processes of inference, cross-situational learning could offer a less taxing means of learning the meaning of words.

While this result only pertains to the limited set of circumstances embodied in our model, it seems to be a promising finding: There is no inherent combinatorial barrier preventing cross-situational learning from scaling up from small lexicons to full-size lexicons under massive referential uncertainty. As such, there is no a priori reason to think that the types of models presented by Yu et al. (2005) and Frank et al. (2009) will necessarily run into difficulty as they move to larger lexicons and increasingly sophisticated corpora.

We now discuss in more detail some of the strengths and weaknesses of the current model, and what modifications could be made to improve it.

The main virtue of this model is that it is sufficiently simple that the central quantity of interest, lexicon learning time, can be calculated exactly. The model reveals that there are two key parameters that quantify the notion of referential uncertainty: the size of the space of meanings that *could* co-occur with the target (M) and the corresponding measure of the number of meanings that *do* co-occur (C). As we have seen, the ratio $\frac{C}{M}$ plays a pivotal role in characterizing the difficulty of the learning task, and hence the lexicon learning time that arises as a result.

Although we have allowed for arbitrary word frequency distributions (citing uniform and Zipfian as two specific examples) and arbitrary overlap between different sets of distractor meanings, it is not clear whether referential uncertainty encountered in reality would be adequately modeled by just two parameters. Even if these two parameters do suffice, their correct values are, at present, unknown. Rather than add further complexity, and with it more unknown parameters to the model, we would advocate determining an empirical estimate of C and M for real-world word learning tasks. For example, the method adopted by Gillette et al. (1999) offers a means of estimating both C and M. Participants in their experiments were presented with short videos of parent–child interactions, with the soundtrack removed and an auditory cue (a beep) inserted to indicate the moment at which the target word is uttered. Participants saw several such videos for each target word, and after viewing each video, participants were asked to make a guess as to the meaning of that word. In our terms, the guesses participants produce after seeing the first video for a given word will tend to be drawn from C. Testing a single video across multiple participants (or asking a single participant to enumerate all possible word meanings for a single video) will therefore offer an indication of the likely membership of C for that usage of the word. Testing across multiple context videos offers some hint as to M for that word: Each video should elicit a different subset C drawn from M. Unlike in our model, we expect that membership of C will be graded, with some frequently guessed members of C and some more marginal members. This would in turn motivate a development of the formal model to include a probabilistic treatment of incidental meanings that allows calculations of lexicon learning time to be made given these more graded notions of context.

We finally discuss some aspects of our model lexicon and learning environment that perhaps oversimplify reality in more serious ways. Despite our lack of knowledge of the true distribution over the set of nontarget meanings (M), it is quite likely that it will not be uniform, as assumed here. Nonuniform distributions will degrade the performance of cross-situational learning relative to one-shot learning, due to the increased likelihood that a frequent nontarget meaning persistently appears whenever a rare target word is uttered. One way to counter this slowdown would be for learners to impose a mutual exclusivity bias (Markman & Wachtel, 1988): an interesting hypothesis to explore would be whether nonuniformity in the environment drives the need for such a bias. Of course, adding constraints like mutual exclusivity to the model would require us to drop the assumption that words are learned independently, which is the simplification that allows us to calculate whole lexicon learning times from single word learning times, perhaps necessitating a different mathematical approach.

We also assume that the lexicon being learned exhibits no ambiguity. Ambiguous words are challenging for a cross-situational learner because, given enough time, a word with two associated meanings will be used in sequences of contexts that have an empty intersection. Siskind (1996) provides a simple but effective work-around that uses empty intersections to identify ambiguous words and repair the lexicon—another (technically challenging) extension to the model would be to calculate how this ambiguity resolution strategy impacts on lexicon learning times for large lexicons.

A related assumption is that of target inclusion: The target meaning is always included in the contexts from which word meaning is inferred. If this assumption is relaxed, an unambiguous word may yield a series of contexts with an empty intersection, due to one or more nonoccurrences of the target meaning—indeed, this is one of the common objections to exclusion-based forms of cross-situational learning (see, e.g, Gleitman, 1990). We note, however, that all theories of word learning must address this issue, and a cross-situational learning strategy that admits large degrees of uncertainty per exposure actually has a robustness advantage compared with approaches that attempt to eliminate uncertainty: Cross-situational learners can include spurious meanings in order to be more sure of including the target meaning and are therefore less likely to eliminate the target erroneously than learners who are less tolerant of referential uncertainty. Cross-situational learning therefore provides a built-in means of dealing with the target elimination problem.[4] Furthermore, weaker variants of cross-situational learning (for example, Approximate XSL) can recover from the occasional nonoccurrence of the target, while still facilitating acquisition of large lexicons in reasonable times.

5. Conclusion

We have shown that cross-situational learning allows the learning of large lexicons in the face of referential uncertainty, at speeds that compare favorably with situations where learners learn individual words more rapidly (e.g., in a single exposure), while potentially offering improved tolerance to noise in the learning environment. Indeed, one could question whether there would be any evolutionary pressure for the powerful heuristics required to drive down referential uncertainty to levels where one-shot word learning routinely becomes possible, given that cross-situational learning offers similar lexicon learning power and requires far weaker constraints. Finally, the techniques we present can be adapted to provide estimates for lexicon learning times for other theories of slow mapping, in order to quantify the link between the speed of individual word learning and the size of the lexicon ultimately attainable. Our calculations suggest that this relationship may be less direct than previously thought: Slow word learning can allow fast learning of large lexicons.

Notes

1. An intriguing alternative possibility, suggested by a reviewer, is that the degree of referential uncertainty experienced by a learner may in part be influenced by the caregiver—for instance, caregivers might manufacture or exploit situations of reduced referential uncertainty in order to facilitate word learning.

2. This assumption also allows us to treat each exposure as involving only a single word: Multiword utterances are simply multiple exposures to single words. While this obviously precludes the explicit inclusion of constraints on word meaning arising from co-occurring words or syntax (as shown to play a key role by, for example, Gleitman, 1990; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005; Gillette, Gleitman, Gleitman, & Lederer, 1999), such constraints can be included in the model in a simplistic fashion as one of the battery of heuristics serving to reduce *C*, in line with our treatment of other heuristics for reducing referential uncertainty.
3. Note that this guess-and-testing learner does not track or make use of the extent of their uncertainty as to a word's meaning—as noted by a reviewer, real-world word learners might be aware of their own uncertainty, which in turn might influence the learning strategy applied.
4. Note that, despite their similarities, this approach can never resolve the problem of homonymous lexical entries, simply because there is no single meaning that is correct for all homonymous words.

Acknowledgments

A. D. M. Smith was funded by Arts and Humanities Research Council grant AR112105 and Economic and Research Council grant RES-062-23-1537. R. A. Blythe is an RCUK Academic Fellow. The authors would like to acknowledge the helpful comments of Paul Vogt, Louise Connell, Simon Dennis, Mike Kalish, Simon Kirby, Dermot Lynott, Franz Mechsner, Catherine O'Hanlon, Elizabeth Wonnacott, and two anonymous reviewers.

References

Abramowitz, M. A., & Stegun, I. A. (1965). *Handbook of mathematical functions*. New York: Dover.

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

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One*, *4*, e7678.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*, 875–890.

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

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

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

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., & Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, *5*, 329–359.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

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

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

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*, 23–64.

Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language*, *21*, 125–155.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27*, 4–9.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*, 128–157.

Jaswal, V. K., & Markman, E. M. (2001). Learning proper and common names in inferential versus ostensive contexts. *Child Development*, *72*, 768–786.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321.

Macnamara, J. (1972). The cognitive basis of language learning in infants. *Psychological Review*, *79*, 1–13.

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

McGregor, K. (2004). Developmental dependencies between lexical semantics and reading. In C. A. Stone, E. R. Silliman, B. J. Ehren, & K. Apel (Eds.), *Handbook of language and literacy* (pp. 302–317). New York: The Guilford Press

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*, 631.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua*, *92*, 377–410.

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

Riley, K. F., Hobson, M. P., & Bence, S. J. (2006). *Mathematical methods for physics and engineering: A comprehensive guide* (3rd ed.). Cambridge, England: Cambridge University Press.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 1–38.

Smith, K., Smith, A. D. M., & Blythe, R. A. (2009). Reconsidering human cross-situational learning capacities: A revision to Yu & Smith's (2007) experimental paradigm. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 2711–2716). Austin, TX: Cognitive Science Society.

Smith, K., Smith, A. D. M., Blythe, R. A., & Vogt, P. (2006). Cross-situational learning: A mathematical approach. In P. Vogt, Y. Sugita, E. Tuci, & C. Nehaniv (Eds.), *Symbol grounding and beyond* (pp. 31–44). Berlin: Springer.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Tomasello, M., & Farrar, J. (1986). Joint attention and early language. *Child Development*, *57*, 1454–1463.

Wilf, H. S. (2006). *Generating functionology*. Wellesley, MA: A. K. Peters.

Wilkinson, K. M., & Mazzitelli, K. (2003). The effect of ''missing'' information on children's retention of fast-mapped labels. *Journal of Child Language*, *30*, 47–73.

Woodward, A. L., & Markman, E. M. (1998). Early word learning. In W. Damon, D. Kuhn, & R. Siegler (Eds.), *Handbook of child psychology, volume 2: Cognition, perception and language* (pp. 371–420). New York: John Wiley and Sons.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied inetntion in early lexical acquisition. *Cognitive Science*, *29*, 961–1005.

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

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Appendix A: Mathematical details

In the main text we claimed that once the way the single-word learning function $P_1(t)$ approaches unity at large times has been identified, the learning time for the whole lexicon $P_W(t)$ can be expressed in terms of the two parameters a and q characterizing this approach and the word frequency distribution ϕ_i; see Eqs. 5 and 7. Here, we justify this claim.

First of all, suppose $P_1(t)$ is known exactly, and that the word with the index i has been exposed $t_i \geq 0$ times. Our central assumption, that all words are learned independently (that is, knowledge of one word's meaning does not improve or diminish the chances of another one being inferred), implies then that, given $P_1(t)$ and the set $\{t_i\}$, the probability all W words have been learned is

$$P_1(t_1)P_2(t_2)\cdots P_W(t_W),$$

no matter what order the exposures have occurred in. We then obtain $P_W(t)$ by summing over all possible $t_1,t_2,\ldots,t_W$ consistent with a total learning time $t=t_1+t_2+\cdots+t_W$. If word i appears with probability ϕ_i in each episode, we find that

$$P_W(t) = \sum_{t_1}\cdots\sum_{t_{W-1}} \frac{t!}{t_1!\cdots t_{W-1}!t_W!}\phi_1^{t_1}P_1(t_1)\cdots\phi_{W-1}^{t_{W-1}}P_1(t_{W-1})\phi_W^{t_W}P_1(t_W) \tag{25}$$

$$= t!\sum_{t_1}\frac{\phi_1^{t_1}}{t_1!}P_1(t_1)\cdots\sum_{t_{W-1}}\frac{\phi_{W-1}^{t_{W-1}}}{t_{W-1}!}P_1(t_{W-1})\frac{\phi_W^{t_W}}{t_W!}P_1(t_W) \tag{26}$$

where the value of t_W is implied by the constraint $\sum_{i=1}^{W} t_i = t$.

The standard way to handle this constraint, and which allows us to approximate this exact expression, is by transforming the functions $P_n(t)$ to their *generating functions* $\mathscr{P}_n(z)$. The key property of a generating function is that it contains the same information as the original function: The coefficients of the tth power of z is equal to $P_n(t)$, so inverting the generating function is a case of reading off the desired coefficient. We will be particularly interested how the coefficients behave as $t \to \infty$, information that can be obtained using a range of analytical techniques (such as Hayman's method) that are described in pedagogical detail in Wilf (2006). We overview the main steps as they apply to the present problem here.

We make use of the exponential generating function that is defined as

$$\mathscr{P}_n(z) = \sum_{t=0}^{\infty}\frac{P_n(t)z^t}{t!}. \tag{27}$$

Then (26) can be expressed equivalently in the extremely compact form

$$\mathscr{P}_W(z) = \prod_{i=1}^{W} \mathscr{P}_1(\phi_i z) \tag{28}$$

which is what allows the different learning strategies to be analyzed.

As $t \to \infty$, we necessarily have that $P_1(t) \to 1$, and hence that to leading order, $\mathscr{P}_1(z) \sim \mathrm{e}^z$. Hence, the leading term in (28) is $\mathrm{e}^{(\sum_i \phi_i)z}$. Since $\sum_i \phi_i = 1$, we find after inverting the generating function that $P_W(t) \to 1$ as $t \to \infty$, as one would expect since $P_1(t) \to 1$ for all W words independently. What is of interest, then, is the *next-leading* term in $\mathscr{P}_W(z)$. This we can read off from the form of $P_1(t)$ common to all the strategies discussed in the main text:

$$P_1(t) = \begin{cases} 0 & t = 0 \\ 1 - a(1-q)^t + r(t) & t > 0 \end{cases} \tag{29}$$

where the remainder term is assumed to have the property that, for some $q' > q$,

$$\lim_{t\to\infty} (1-q')^t r(t) = 0. \tag{30}$$

In the following, it is useful to keep in mind the largest value of $\Delta = q' - q$ for which this limit holds: This gives an indication of when the next-next-leading term becomes relevant, and the approximation that $P_1(t)$ is completely characterized by the two parameters a and q breaks down.

Given these definitions, we find that

$$\mathscr{P}_1(z) \sim \mathrm{e}^z \left[1 - a\mathrm{e}^{-qz} + O(\mathrm{e}^{-(q+\Delta)z})\right]. \tag{31}$$

Evaluating now the saddle-point (Riley, Hobson, & Bence, 2006) of the inversion integral (which is what is involved in the application of Hayman's method; Wilf, 2006),

$$P_W(t) = \frac{1}{2\pi i} \oint \frac{\mathrm{d}z}{z^{t+1}} \mathscr{P}_W(t), \tag{32}$$

we ultimately find that

$$P_W(t) = \prod_{i=1}^{W} [1 - a\mathrm{e}^{-q\phi_i t} + O(\mathrm{e}^{-(q+\Delta)\phi_i t})]. \tag{33}$$

Truncating each multiplicand after the second term—which is what is done to arrive at (7)—is valid if $\Delta\phi_i t^* \gg 1$ for any i; hence, we arrive at the criterion $t^* \gg 1/(\phi_{\min}\Delta)$ for the validity of the learning time t^* obtained from (7).

For the fast mapping strategy, this truncation involves no approximation (Δ is effectively infinite in this case). For the Minimal XSL strategy, $\Delta = \frac{C}{C+1}\frac{M+1}{M}$ and for the Pure XSL strategy, $\Delta = \frac{C}{M-1}(1 - \frac{C}{M})$. We thus find that the result for the Minimal XSL strategy given in the main text holds if C exceeds $[\,\ln W/\epsilon]^{-1}$, which is always true if W/ϵ is larger than about 3 (and therefore certainly valid when W is large and ϵ small, which is the range of interest). Meanwhile, the result for Pure XSL holds if C/M is larger than $[\ \ln W/\epsilon]^{-1}$. For the values of $W = 60{,}000$, $\epsilon = 0.01$, and $M = 100$ used in the main text, this corresponds to C being larger than about 6. We see in Fig. 2 that the theoretical prediction does indeed differ from the values obtained from the Monte Carlo simulation in this regime. Note that these conditions hold both for the uniform and the Zipfian distributions.

Appendix B: Monte Carlo methods

Since a number of approximations were made in deriving the learning time formulæ, it is worthwhile to compare these predictions with data obtained from Monte Carlo simulations of the model learning tasks discussed. Furthermore, in the absence of analytical predictions for the Approximate XSL strategy described in the main text, simulation is the only means we have at our disposal to obtain the requisite data for Fig. 3.

In principle, the simulation proceeds as follows. A random number generator (specifically a Mersenne twister) is used to generate a sequence of target meanings, drawn at random from the set of W available targets according to the appropriate distribution (uniform or Zipf). In each of these episodes, C distinct nontarget meanings are also selected from the M possibilities. If the target meaning has never been presented before, one of the $C + 1$ meanings present is chosen at random as the current hypothesis for that meaning. If this hypothesis is correct, the word is marked as learned, as all further exposures will confirm the correct hypothesis. On subsequent exposures of unlearned words, the hypothesis is retained if it coincides with one of the meanings present, or a new hypothesis is chosen either uniformly from the scene (Minimal XSL), frequency-weighted from the scene (Approximate XSL), or from the set of confounding meanings (XSL). In the Approximate XSL case, it is necessary to keep track of the number of times each meaning has appeared alongside a given target, and in the full XSL case the set of confounding meanings must be tracked. The simulation stops when all words have been learned and the number of episodes needed to reach that point is output. To obtain the learning times shown in the figures, a sample of $N = 2{,}000$ learning times was generated for each, and the time $t^*(\epsilon)$ obtained by dividing this sample into two sets, one containing the largest $N\epsilon(= 20$ for $\epsilon = 0.01)$ learning times, and the other containing the rest. The numerical value of $t^*(\epsilon)$ was then taken to be the midway point between the smallest element of the former set and the largest of the latter.

In practice, a more optimized version of the above was actually used to generate the data shown in Figs. 2 and 3. For example, in the Minimal XSL case we can notionally maintain all possible false hypotheses in parallel, switching with probability $1 - C/M$ in each episode whereupon a correct hypothesis is then chosen with probability $1/(C + 1)$. Each possible learning time is still generated with the desired probability, but this approach allows for better statistics from fewer samples. A similar optimization was employed in the other two cases.

In all cases we found the Monte Carlo results to be in excellent agreement with the theoretical predictions where the latter were available. The only exception to this is for the full XSL strategy in the small C/M regime for the reasons we have discussed above.