Cross-Situational Learning: A Mathematical Approach

Kenny Smith[1], Andrew D.M. Smith[1], Richard A. Blythe[2], and Paul Vogt[1,3]

[1] Language Evolution and Computation Research Unit, University of Edinburgh
[2] School of Physics, University of Edinburgh
[3] ILK/Language and Information Science, Tilburg University, The Netherlands
{kenny, andrew, paulv}@ling.ed.ac.uk, R.A.Blythe@ed.ac.uk

Abstract. We present a mathematical model of cross-situational learning, in which we quantify the learnability of words and vocabularies. We find that high levels of uncertainty are not an impediment to learning single words or whole vocabulary systems, as long as the level of uncertainty is somewhat lower than the total number of meanings in the system. We further note that even large vocabularies are learnable through cross-situational learning.

1 Introduction

One of the design features of human language is the *arbitrary* relationship between words and their meanings [1] — they are not related iconically, through perceptual similarity, but merely by convention. Learning word-meaning mappings is therefore far from trivial, yet when children acquire language, they learn the meanings of a large number of words very quickly. This phenomenon is known as *fast mapping* [2]. Precisely how children achieve this remains to be established.

The problem of *referential indeterminacy* in acquiring word–meaning mappings was famously illustrated by Quine [3]. 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 IT WILL RAIN.

Developmental linguists have proposed many mechanisms which children may use to overcome referential indeterminacy in word learning (see [4,5] for overviews). Tomasello, for instance, proposes that the core mechanism is *joint attention* [6,7]; children understand that adults use utterances to refer to things, and upon hearing an utterance they attempt to attend to the same situation as their caregivers. Establishing joint attention in this way reduces the number of potential meanings a word might have, although Quine shows that this cannot be sufficient. Researchers have proposed a number of representational biases (e.g. the *whole object bias* [8] and the *shape bias* [9]) and interpretational constraints (e.g. *mutual exclusivity* [10] and the *principle of contrast* [11]) which might act to further reduce the indeterminacy problem.

P. Vogt et al. (Eds.): EELC 2006, LNAI 4211, pp. 31–44, 2006.

Further evidence suggests that children may learn the meaning of many words more straightforwardly, by simply disambiguating potential meanings across different occasions of use [12,13]. There is evidence that this process, known as *cross-situational learning*, takes place from a very early age [14]. Cross-situational learning is unlikely to provide a complete account of word learning, but does allow us to consider word learning in the absence of sophisticated cognitive mechanisms.

Understanding how children learn the meaning of words is not only a key question in developmental linguistics, but is also fundamentally an evolutionary issue. Firstly, accounting for the design feature of arbitrariness requires us to understand how the apparent problems introduced by arbitrary meaning-word mappings might be resolved. Secondly, an account of the evolution of the capacity for language must begin with a clear specification of the explanandum — for example, must the capacity for language include domain-specific word learning strategies? Finally, the indeterminacy of meaning is itself a important issue in the literature on the computational modelling of linguistic evolution [15,16]

In this paper, we present a mathematical model of cross-situational language learning and use it to quantify some basic properties of the learnability of words and vocabularies. In the following section, we describe cross-situational learning in more detail. Our formalisation is introduced in section 3, where we quantify the learnability of individual utterances. In section 4, we extend the model to quantify the learnability of a whole language. Finally, in section 5 we discuss the study's implications, and explore extensions of the model to address more realistic treatments of language structure, use and learning.

2 Cross-Situational Learning

Cross-situational learning is a technique for working out the reference of an utterance, based on multiple exposures to the utterance's use in context. When an utterance is produced, the context of its use will provide a number of candidate meanings for that utterance. From a hearer's point of view, each of these is in principle equally plausible, and there is no obvious motivation for choosing between them. If the same utterance is produced in a different situation, however, a different set of possible meanings may be suggested by that situation. The hearer can make use of this, by taking the intersection of the two sets of possible meanings, in order to (potentially) reduce the ambiguity of the utterance.

Cross-situational learning has been modelled computationally by Siskind [17], who showed that it could indeed be used to learn word-meaning mappings. In his model, a learner is exposed to a corpus of artificial sentences, each of which is paired with a set of possible meanings. Initially, the learner associates each word with all possible meanings. When hearing a word in a new situation, however, the learner eliminates any existing meanings for that word which are not consistent with the new situation.

Variants of the cross-situational model have been used to simulate the evolution of lexicons in multi-agent systems [16,18], in which meanings are built up through interaction with the world and other individuals. In these experiments, Smith [18] and Vogt [16] have separately shown that conventionalised vocabularies can emerge and persist through cross-situational learning. Our focus in this paper is similar to Siskind's — we are interested in the learnability of an existing vocabulary system, rather than the negotiation of shared vocabularies in a population. However, our approach is different — rather than modelling cross-situational learning computationally, we seek as far as possible an exact mathematical characterisation of the properties of the system. This paper represents a preliminary stage in this process.

3 The Mathematical Model of Cross-Situational Learning

In this section, we describe a mathematical model which we can use to specify the probability of a learner learning the meaning of a word cross-situationally. In every episode of exposure to an utterance, the hearer observes a situation which provides both the intended meaning of the utterance (the *target* meaning) and a set of other meanings incidentally provided by the situation (the *context*).

Assume that the context has the same number of members C in each episode, but the members are chosen at random and without duplication from the larger set of M possible meanings.[1] There are therefore $\binom{M}{C}$ different possible contexts.

Let the context in episode E_e be C_e. If, after e episodes, a non-target meaning has occurred in every episode $E_1 \ldots E_e$, then that meaning is called a *confounder* — this recurring meaning is an equally plausible meaning for the utterance as the target meaning, given that it too is present in all e situations where the utterance is used. Let the number of confounders after e episodes be K_e, and let us assume that the meaning of a word is successfully learned after e episodes if there are no confounders left ($K_e = 0$) — when $K_e = 0$, the target meaning is the only one which has occurred in every one of the e episodes.

3.1 An Illustration

Let us take a simple example, with $C = 3$ and $M = 5$. The 10 possible contexts are enumerated in Fig. 1, and we assume for this exposition that they are equiprobable, and that each therefore occurs with a probability of $\binom{M}{C}^{-1}$. In the graphical notation in Fig. 1, each context is represented as a row of M boxes, with each box representing a meaning. A cross in a box denotes that that meaning is present in the given context.

Note that there are necessarily C confounders ($K_1 = C$) after E_1 — each of the meanings in context C_1 has occurred as often as the target meaning, namely once. Let us now investigate what happens in episode E_2, taking context $E_1 = \{m_1, m_2, m_3\}$ as an example, and combining it with each possible context which

[1] Note that M is *exclusive* of the target meaning. In other words, there are $M + 1$ possible meanings, and any situation provides $C + 1$ unique meanings: the target and C unique additional meanings.

$m_1 m_2 m_3 m_4 m_5$	$m_1 m_2 m_3 m_4 m_5$	$m_1 m_2 m_3 m_4 m_5$	$m_1 m_2 m_3 m_4 m_5$	$m_1 m_2 m_3 m_4 m_5$
☒☒☒□□	☒☒□☒□	☒☒□□☒	☒□☒☒□	☒□☒□☒
$m_1 m_2 m_3 m_4 m_5$	$m_1 m_2 m_3 m_4 m_5$	$m_1 m_2 m_3 m_4 m_5$	$m_1 m_2 m_3 m_4 m_5$	$m_1 m_2 m_3 m_4 m_5$
☒□□☒☒	□☒☒☒□	□☒☒□☒	□☒□☒☒	□□☒☒☒

Fig. 1. Enumeration of $\binom{M}{C} = 10$ possible contexts, with $C = 3$ and $M = 5$

could occur in episode E_2. Fig. 2 below shows the 10 resultant combinations, the number of confounders K_2, and the confounder meanings highlighted in grey.

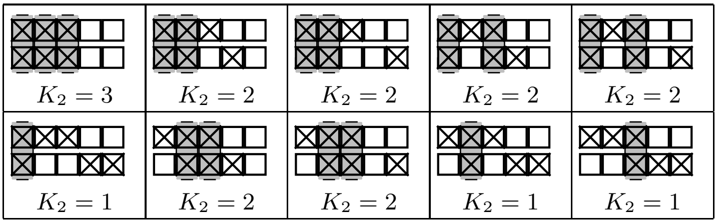

Fig. 2. Combinations of contexts after E_2, with the number of confounders K_2, and the confounder meanings highlighted in grey

We can see in Fig. 2 that the set of confounders remaining after episode E_2 is dependent on the set of confounders from E_1, and the meanings in C_2. We can ignore all meanings which did not occur in C_1, as they can never be confounders — a single non-occurrence in one episode is enough to rule out a particular meaning as a confounder. More generally, the set of confounders K_e after episode E_e depends on the set of confounders after the previous episode E_{e-1} , namely K_{e-1}, and the set of meanings chosen in context C_e.

Let the probability of having n confounders after e episodes $P(K_e = n)$ be $P_n(e)$. The probability that a word is successfully learned after e episodes is therefore $P_0(e)$. After E_2, and assuming $C_1 = \{m_1, m_2, m_3\}$, we can see in Fig. 2 that $P_3(2) = \frac{1}{10}$; $P_2(2) = \frac{6}{10}$; $P_1(2) = \frac{3}{10}$ and $P_0(2) = \frac{0}{10}$. Note in this case that it is impossible to have learned a word after two episodes ($P_0(2) = 0$), because the context is larger than half of the number of possible meanings ($C > \frac{M}{2}$), and so it is impossible to select disjoint sets for C_1 and C_2. It should be clear that the choice of $C_1 = \{m_1, m_2, m_3\}$ in this example is unimportant: the same probabilities for each value of K_2 are obtained for every possible choice for C_1.

What happens, however, when there are fewer than C confounders at the previous timestep ($K_{e-1} < C$)? To examine this situation we have to look at a further episode, E_3. Let's take $C_1 = \{m_1, m_2, m_3\}, C_2 = \{m_1, m_2, m_4\}$ as an example, giving $K_2 = 2$, and combine it with all possibilities for C_3, as depicted in Fig. 3.

We can see that for $K_2 = 2$, given $C_1 = \{m_1, m_2, m_3\}$ and $C_2 = \{m_1, m_2, m_4\}$, the probabilities are $P_2(3) = \frac{3}{10}, P_1(3) = \frac{6}{10}, P_0(3) = \frac{1}{10}$. The choice of C_1 and C_2

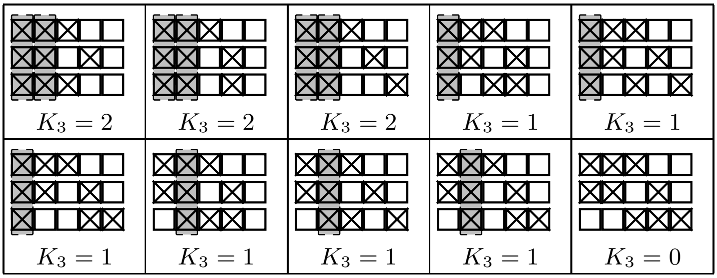

Fig. 3. Combinations of contexts after E_3,with the number of confounders (K_3), and the confounder meanings highlighted in grey

is again unimportant, as the same probabilities for each value of K_3 are obtained for each combination where $K_2 = 2$. Similar calculations can be carried out for $K_2 = 1$, by choosing (for instance) $C_1 = \{m_1, m_2, m_3\}$ and $C_2 = \{m_1, m_4, m_5\}$.

3.2 Calculating Semantic Inferrability

In general, the transition probability $Q(x|y)$, i.e. that there will be x confounders after episode e, given that there were y confounders after episode $e-1$, is:

$$Q(x|y) = \binom{y}{x} \times \binom{M-y}{C-x} \times \binom{M}{C}^{-1} \tag{1}$$

The first term $\binom{y}{x}$ is the number of ways of correctly selecting *confounders*: y is the number of confounders at time $e-1$ (call this the confounding set), and x is the number of confounders we want to have at time e. There are therefore $\binom{y}{x}$ ways in which the desired number of confounders x can be chosen from the confounding set y. The second term $\binom{M-y}{C-x}$ is likewise the number of ways of correctly selecting non-confounders: $M-y$ gives the number of meanings which are *not* confounders at time $e-1$ (call this the non-confounding set). Recall that every context has C members, so if there are x confounders in a valid context, then we must also select $C-x$ non-confounders from the non-confounding set. There are clearly $\binom{M-y}{C-x}$ ways of choosing the desired number of non-confounders $C-x$ from the non-confounding set $M-y$, as shown in Fig. 4. The number of valid contexts which satisfy the desired condition is the product of these two expressions, divided by the total number of possible contexts, to produce the overall transition probability Q.

Therefore, the probability $P_n(e)$, that there will be n confounders after e episodes is:

$$P_n(e) = \sum_{i=n}^{C} P_i(e-1) \times Q(n|i) \; . \tag{2}$$

We have already seen, however, that if $e = 1$, then the number of confounders is necessarily C, so for completeness (2) should be extended to cover the case

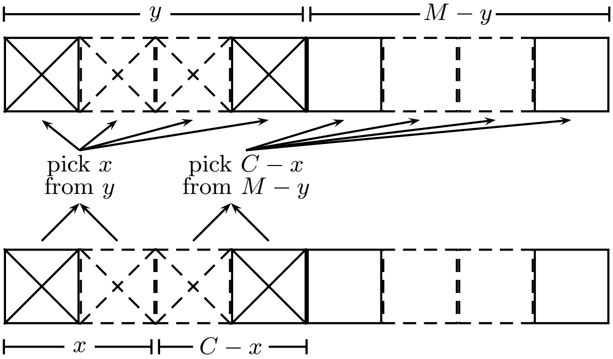

Fig. 4. Building a context of size C, made up of x confounders chosen from the y members of the confounding set, and $C - x$ non-confounders chosen from the $M - y$ members of the non-confounding set.

where $e = 1$:

$$P_n(e) = \begin{cases} 1 & \text{if } e = 1,\, n = C, \\ 0 & \text{if } e = 1,\, n \neq C, \\ \sum_{i=n}^{C} P_i(e-1) \times Q(n|i) & \text{otherwise.} \end{cases} \tag{3}$$

In Appendix A, we solve (3) to give the following explicit formula for $P_n(e)$:

$$P_n(e) = \binom{C}{n} \sum_{i=n}^{C} (-1)^{i-n} \binom{C-n}{i-n} (p_i)^{e-1} \tag{4}$$

where

$$p_i = \frac{\binom{M-i}{C-i}}{\binom{M}{C}} = \begin{cases} 1 & \text{for } i = 0 \\ \frac{C(C-1)\ldots(C-i+1)}{M(M-1)\ldots(M-i+1)} & \text{for } i > 0 \end{cases} \tag{5}$$

is the probability that a particular subset of i members of the C confounders in the first episode E_1 appear in any subsequent episode.

3.3 Word Learnability Results

Using either (3) or (4), therefore, we can quantify the learnability of an *individual* word — the probability that an individual word will be learned, $P_0(e)$ — which depends on M, C, and e. Fig. 5 shows word learnability for $M = 50$, for various values of C. Two basic results are apparent: (i) A word cannot be learned when $C = M$, as confounders can never be eliminated; (ii) For all other cases, learnability increases over time, although it may be the case (for example, when C is high) that learnability remains at zero for a number of exposures, before becoming non-zero.

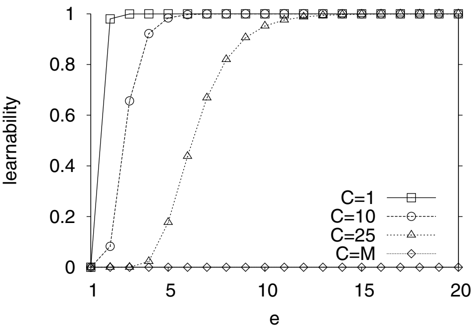

Fig. 5. Word learnability given $M = 50$, for various C

We can also quantify the number of episodes e^* required to learn a word with probability $1 - \epsilon$. Fig. 6 (a) shows e^* given $M = 50$, with $\epsilon = 0.01$, for various context sizes. Expected values are derived from Eqn. (3), exact values by Monte Carlo simulation[2]. It is clear that the results from the Monte Carlo simulation closely match the results from the mathematical model. In addition, we see that (iii) the smaller the context size, the quicker a word can be learned; (iv) as C approaches M, it takes a long time to learn a word, as confounders are only rarely eliminated. Fig. 6 (b) shows e^* given $C = 5$, with $\epsilon = 0.01$, for various M. We can see that (v) words can be learned more rapidly as the number of meanings increases; as M increases, it becomes less likely that any one meaning will recur in every context with the target meaning.

4 Quantifying the Learnability of a Whole Language

The model described in the previous section only considers the learnability of a *single word*. One conclusion is that, given a fixed context size, the meaning of a particular word is easier to learn if that word is part of a large system for conveying a large number of distinct meanings (M is large). This suggests that we need to consider the learnability of a whole vocabulary system consisting of a number of words, each of which conveys a particular meaning, rather than considering word learnability in isolation.

In order to do this, we must first introduce a minor change to our notation. When considering the learnability of a single word, we were concerned with the number of meanings *other than the target meaning*, and the number of meanings in the context *other than the target meaning*. We denoted these as M and C respectively. When quantifying the learnability of a whole set of words, we are necessarily interested in cases where the target meaning for a particular word may also occur as a non-target meaning for some usage of some other word. Let

[2] In the simulation, a learner works through a series of exposures, eliminating candidate meanings. e^* is the number of episodes required to achieve learnability of $1 - \epsilon$ averaged over 1000 such simulations.

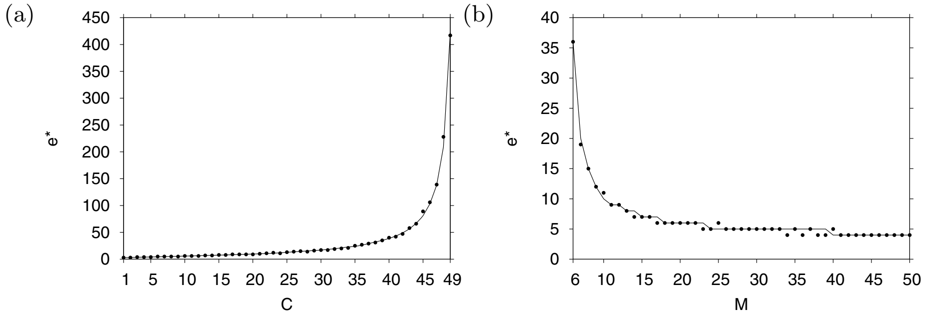

Fig. 6. The number of episodes required to learn a word with probability 0.99 varies with the number of meanings and the context size; (a) shows e^* given $M = 50$, for various C, (b) shows e^* given $C = 5$, for various M. Lines are expected values, points are actual (Monte Carlo simulation) values.

us therefore call the total number of lexicalised meanings in a vocabulary system $\bar{M}$. In every episode of exposure to an utterance conveying one of these meanings, the hearer observes a situation which provides both the target meaning and a context of other meanings. The number of meanings involved in the context, *inclusive of the target meaning*, is given by $\bar{C}$. The $C = \bar{C} - 1$ non-target meanings in the context are chosen at random and without duplication from the larger set of $M = \bar{M} - 1$ possible meanings. In other words, $\bar{M}$ and $\bar{C}$ are inclusive, rather than exclusive, of the target meaning.

It is convenient, at least initially, to consider the situation where only W of the total number of possible meanings $\bar{M}$ are ever chosen as the target. We seek now $R_W(e)$, the probability that all W of these words have been learned after e episodes; the probability that the whole language has been learned is then given by the special case $W = \bar{M}$. To obtain this, we must average over all W^e sequences of utterances. Some particular sequences may, or may not, be equivalent to one another depending on what inferences are made by the learner. If, for example, the learner assumes that different words do not have the same meaning, then the order with which the words are presented matters. Under this assumption, if the word for a meaning is learned then that meaning can no longer act as a confounder for the remaining meanings. This induces non-trivial interactions between episodes in which different words are uttered. On the other hand, if the learner entertains the possibility that two words may have the same meaning, then they must wait until all meanings other than the target have been ruled out. In this latter case, the probability that a meaning has been learned is independent of the order in which the words are presented, and thus depends only on the number of times a particular meaning has been chosen as the target. In this much simpler case, which we will focus on here, only the number of times a word is uttered is important: order of presentation does not matter.

In this case, the probability of learning all W words is given by

$$R_W(e) = \langle P_0(e_1) P_0(e_2) \cdots P_0(e_W) \rangle \tag{6}$$

where the angle brackets denote an average over the probability distribution of sequences of e episodes in which the first word of interest is the target e_1 times, the second e_2 and so on. This distribution is the multinomial distribution

$$\frac{1}{W^e}\binom{e}{e_1 e_2 \cdots e_W} \equiv \frac{1}{W^e}\frac{e!}{e_1!e_2!\cdots e_W!}$$

constrained such that $\sum_i e_i = e$. The functions $P_0(e_i)$ appearing in Eqn. (6) are as given by Eqn. (4). It is possible to calculate this average exactly; unfortunately, the expression that results is rather unwieldy and extremely difficult to interpret. We thus derive instead an approximation to $R_W(e)$ that admits a clearer insight into the learnability of an entire language.

This approximation is obtained by focusing on the regime where the language is learnt to a high probability, i.e., where $R_W(e) = 1 - \epsilon_W$ and the parameter ϵ_W is small. For example $\epsilon_W = 0.01$ corresponds to having learned the words with 99% certainty. In Appendix B, we present the details of this approximate approach which results in the following expression for the probability of learning W of $\bar{M}$ words after e episodes:

$$R_W(e) \approx \sum_{k=0}^{W} \binom{W}{k} (1-\bar{M})^k \left[1 - \frac{k}{W}\left(\frac{\bar{M}-\bar{C}}{\bar{M}-1}\right)\right]^e . \tag{7}$$

Since each term in the series is progressively smaller, and the relative size of each term is roughly equal to the absolute size of the previous term, the series can be truncated at $k = 1$ as long as ϵ_W is sufficiently small. Inverting this truncated expression gives an indication of the time taken to learn the whole language with probability $1 - \epsilon_W$. It reads

$$e^* \approx \frac{\ln[\epsilon_W] - \ln[W(\bar{M}-1)]}{\ln\left[1 - \frac{1}{W}\left(\frac{\bar{M}-\bar{C}}{\bar{M}-1}\right)\right]} . \tag{8}$$

Since various approximations have been made to arrive at this formula, it is worth testing its validity by comparing with data from Monte Carlo simulations. Fig. 7 shows the match between expected and actual (obtained from simulation) values given various values of ϵ, $\bar{C}$ and $\bar{M} = W$. As can be seen from the figures, there is close agreement between the actual and expected values as long as ϵ_W is not large (Fig. 7 (a)) and $\bar{C}$ is not small (Fig. 7 (b)). The former condition is easily understood, since ϵ_W was assumed to be small throughout the derivation of (7) and (8). Meanwhile, a closer analysis of the approximations used in Appendix B to derive these expressions shows that strong fluctuations in the number of episodes required to learn a single word lead to the breakdown of the approximation when $\bar{C}$ is small.

Fig. 7 (b) shows e^* given $M = 50$, $\epsilon_W = 0.01$, for various context sizes. It is apparent that (i) the smaller the context size, the quicker a whole vocabulary can be learned; (ii) as $\bar{C}$ approaches $\bar{M}$, it takes a long time to learn a word, as confounders are only rarely eliminated. In other words, $\bar{C}$ does not have to

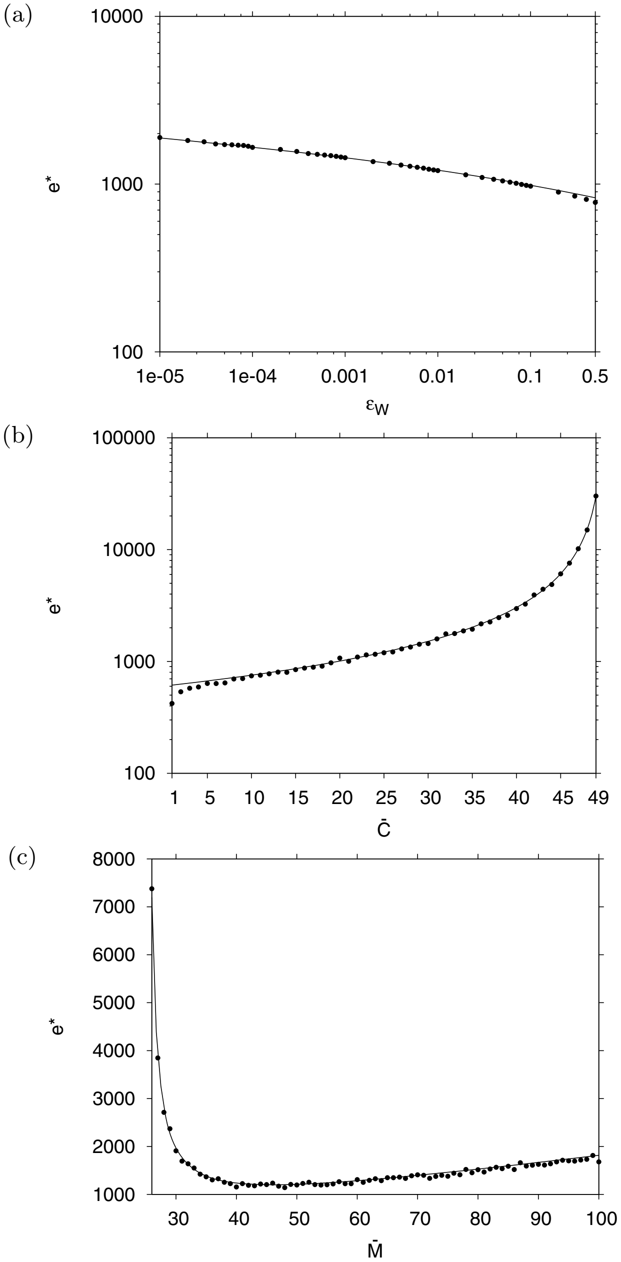

Fig. 7. The number of episodes needed to learn a whole vocabulary with probability $1 - \epsilon_W$. (a) shows e^* given $\bar{M} = 50, \bar{C} = 25$, for various ϵ_W. (b) shows e^* given $\bar{M} = 50, \epsilon_W = 0.01$, for various $\bar{C}$. (c) shows e^* given $\bar{C} = 25, \epsilon_W = 0.01$, for various $\bar{M}$. Lines are expected values, points are actual (Monte Carlo) values. Note log scales on (a) and (b).

be very small for a vocabulary to be learned in a reasonable time, as long as it is fairly small *relative to* $\bar{M}$. Fig. 7 (c) shows e^* given $\bar{C} = 25, \epsilon_W = 0.01$, for various $\bar{M}$. Here we see that (iii) it is easiest to learn a whole language when $\bar{C}$ is less than $\bar{M}$ and both are relatively small.

Fig. 7 (c) further suggests that, once $\bar{M}$ is significantly greater than $\bar{C}$, e^* increases linearly with $\bar{M}$. In fact, putting $W = \bar{M}$ in Eqn. (8) suggests that the rate of increase is slightly greater than linear. Specifically, one finds that once $\bar{M}$ has greatly exceeded the larger of $\bar{C}$ and $\ln \epsilon_W$,

$$e^* \sim 2\bar{M} \ln \bar{M} \ . \tag{9}$$

In other words, (iv) while the time taken to learn a vocabulary of a particular size increases superlinearly with respect to the size of that vocabulary, there is no critical value of $\bar{M}$ beyond which e^* increases dramatically — large vocabularies are learnable through cross-situational learning.

5 Discussion

We have outlined a mathematical formulation of cross-situational learning, and presented some basic results linking word and vocabulary learnability to the size of the vocabulary system, the number of candidate meanings provided by a context of use, and the amount of time for learning. Based on these results, it is tempting to speculate on the human case, particularly from an evolutionary perspective — for example, we might claim that humans have a long period of developmental flexibility to allow them time to learn a large vocabulary system, or that humans have evolved a number of biases for word-learning to reduce the effective context size during word learning and make large vocabularies learnable in a fairly short period of time.

However, several shortcomings in the model as it stands need to be addressed before such speculations can be entertained (if at all). Firstly, and most importantly, we have considered both words and meanings to be unstructured atomic entities. The model as it stands is therefore better interpreted as quantifying the learnability of a *holistic* system. In *compositional* systems, such as language, meanings are structured objects and utterances are structured sequences of words. We are currently extending this model to explore such a situation, in order to contrast the learnability of words in systems of different structural kinds.

Secondly, we assume that all meanings occur with uniform probability. This is unlikely to be exactly true, and it may be that the frequency of communicatively-relevant situations is highly non-uniform, possibly Zipfian [19]. How does this affect word learnability? Again, we are extending our model to allow us to investigate such questions.

Finally, as discussed in section 4, we have assumed that the meaning of each word is learned independently — learning something about the meaning of one word tells you nothing about the meaning of another word. We know, however, that this assumption is not true for humans, who instead appear to assume that if one word has a particular meaning, then no other word will have that same meaning — this is mutual exclusivity [10]. How much, if at all, does mutual exclusivity simplify the learning of words in holistic or structured systems? We are also investigating this question using a Monte Carlo version of our model.

The model outlined here is, we feel, an important first step on the path to a more thorough and formal understanding of the developmental and evolutionary problem of word learning.

References

1. Hockett, C.F.: The origin of speech. Scientific American **203** (1960) 88–96
2. Carey, S., Bartlett, E.: Acquiring a single new word. Papers and Reports on Child Language Development **15** (1978) 17–29
3. Quine, W.v.O.: Word and Object. MIT Press, Cambridge, MA (1960)
4. Bloom, P.: How Children Learn the Meanings of Words. MIT Press, Cambridge, MA (2000)
5. Hall, D.G., Waxman, S.R., eds.: Weaving a Lexicon. MIT Press, Cambridge, MA (2004)
6. Tomasello, M.: The cultural origins of human cognition. Harvard University Press, Harvard (1999)
7. Tomasello, M.: Constructing a language: a usage-based theory of language acquisition. Harvard University Press (2003)
8. Macnamara, J.: Names for things: a study of human learning. MIT Press, Cambridge, MA (1982)
9. Landau, B., Smith, L.B., Jones, S.S.: The importance of shape in early lexical learning. Cognitive Development **3** (1988) 299–321
10. Markman, E.M.: Categorization and naming in children: problems of induction. Learning, Development and Conceptual Change. MIT Press, Cambridge. MA (1989)
11. Clark, E.V.: The lexicon in acquisition. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge (1993)
12. Akhtar, N., Montague, L.: Early lexical acquisition: the role of cross-situational learning. First Language (1999) 347–358
13. Klibanoff, R.S., Waxman, S.R.: Basic level object categories support the acquisition of novel adjectives: evidence from pre-school aged children. Child Development **71 (3)** (2000) 649–659
14. Houston-Price, C., Plunkett, K., Harris, P.: 'Word-Learning Wizardry' at 1;6. Journal of Child Language **32(1)** (2005) 175–189
15. Smith, A.D.M.: Establishing communication systems without explicit meaning transmission. In Kelemen, J., Sosík, P., eds.: Advances in Artificial Life. Springer-Verlag, Heidelberg (2001) 381–390
16. Vogt, P., Coumans, H.: Investigating social interaction strategies for bootstrapping lexicon development. Journal of Artificial Societies and Social Simulation **6**(1) (2003) http://jasss.soc.surrey.ac.uk/6/1/4.html.
17. Siskind, J.M.: A computational study of cross-situational techniques for learning word-to-meaning mappings. Cognition **61** (1996) 39–91
18. Smith, A.D.M.: Intelligent meaning creation in a clumpy world helps communication. Artificial Life **9**(2) (2003) 175–190
19. Zipf, G.K.: The Psycho-Biology of Language. Routledge, London (1936)
20. Wilf, H.S.: Generatingfunctionology. Academic Press (1994)

A Exact Solution for the Single Word Case

The exact solution given in Eqn. (4) can be obtained in two ways: (i) by diagonalisation of the matrix of transition probabilities $Q(x|y)$; or (ii) by applying the "inclusion-exclusion" principle (or sieve method) from combinatorics. In this Appendix, we outline the latter approach which, as explained by Wilf [20, p.110], is useful when "it is relatively easy to see how many objects have at least a certain number of properties and maybe more". The sieve method, he goes on to explain, converts this "at least" information into the desired "exactly" information.

In our application, we seek $P_n(e)$, the probability that n of the initial C confounders are present in each of a number e of episodes. The "at least" information here is the probability p_n that a specific subset of n confounders appears in each of e episodes, along with maybe some other confounders. This probability is given by p_n^{e-1} Eqn. (5), since the desired subset is always present in the first episode (by definition), and then with probability p_n in subsequent episodes.

The sieve method then states that the probability of having a subset of N confounders present in every episode is given by the sum

$$P_n(e) = \sum_{i=n}^{C} (-1)^{i-n} \binom{i}{n} \sum_{i\text{-subsets of } C \text{ confounders}} p_i^{e-1} \qquad (10)$$

$$= \sum_{i=n}^{C} (-1)^{i-n} \binom{i}{n} \binom{C}{i} p_i^{e-i} \qquad (11)$$

where we have used the fact that there are $\binom{C}{i}$ distinct subsets of size i contained within a set of C objects. The result (4) then follows from the fact that $\binom{i}{n}\binom{C}{i} = \binom{C}{n}\binom{C-i}{i-n}$, as can be verified by writing the binomial coefficients explicitly in terms of factorials.

B Approximate Solution for the Multiple Word Case

We are interested in determining the probability $R_W(e)$ that W of $\bar{M}$ meanings have been learnt after a total number of e episodes in the regime where $R_W(e) \approx 1$. Our approach rests on the following observation: if all W words are to be learnt with certainty $1-\epsilon_W$ (ϵ_W being a small parameter), each of the factors $P_0(e_i)$ in Eqn. (6) should contribute an amount approximately equal to $1 - \frac{\epsilon_W}{W}$. That is, every word has to be learnt (on average) to a *higher* level of certainty; the value of ϵ for a single word (ϵ_1) is approximately equal to $\frac{\epsilon_W}{W}$. Looking at Fig. 5, we see that to achieve this high level of single-word learnability, many utterances of each individual word are required in order to eliminate all confounding meanings. The upshot of this is that, since e_i is expected to be large, the expression for $P_0(e_i)$, Eqn. (4), is well approximated by the first two terms in the series. We henceforth assume that we can write

$$P_0(e_i) \approx 1 - (\bar{M} - 1) \left(\frac{\bar{C} - 1}{\bar{M} - 1} \right)^{e_i} . \qquad (12)$$

Using this approximation in Eqn. (6) we find

$$R_W(e) \approx \left\langle \prod_{i=1}^{W} \left[1 - (\bar{M} - 1) \left(\frac{\bar{C} - 1}{\bar{M} - 1} \right)^{e_i} \right] \right\rangle \tag{13}$$

$$= \sum_{k=0}^{W} \binom{W}{k} (1 - \bar{M})^k \left\langle \left(\frac{\bar{C} - 1}{\bar{M} - 1} \right)^{e_1 + e_2 + \cdots + e_k} \right\rangle . \tag{14}$$

The average over the multinomial distribution can then be computed by noting the identity

$$\sum_{e_1} \sum_{e_2} \cdots \sum_{e_W} \binom{e}{e_1 e_2 \cdots e_W} u_1^{e_1} u_2^{e_2} \cdots u_W^{e_W} = (u_1 + u_2 + \cdots + u_W)^e \tag{15}$$

which yields Eqn. (7). As we note in the text, the approximation (12) holds as long as fluctuations in the number of episodes in which a particular meaning is the target are small relative to the mean.