Edinburgh Research Explorer

Word learning under infinite uncertainty

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
Cognition

OPEN ACCESS

Word learning under infinite uncertainty

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

[a]*SUPA, School of Physics and Astronomy, University of Edinburgh, Edinburgh EH9 3JZ, UK*
[b]*Literature and Languages, School of Arts and Humanities, University of Stirling, Stirling, FK9 4LA, UK*
[c]*School of Philosophy, Psychology and Language Sciences, University of Edinburgh, EH8 9AD, UK*

Abstract

Language learners must learn the meanings of many thousands of words, despite those words occurring in complex environments in which infinitely many meanings might be inferred by the learner as a word's true meaning. This problem of *infinite referential uncertainty* is often attributed to Willard Van Orman Quine. We provide a mathematical formalisation of an ideal cross-situational learner attempting to learn under infinite referential uncertainty, and identify conditions under which word learning is possible. As Quine's intuitions suggest, learning under infinite uncertainty is in fact possible, provided that learners have some means of ranking candidate word meanings in terms of their plausibility; furthermore, our analysis shows that this ranking could in fact be exceedingly weak, implying that constraints which allow learners to infer the plausibility of candidate word meanings could themselves be weak. This approach lifts the burden of explanation from 'smart' word learning constraints in learners, and suggests a programme of research into weak, unreliable, probabilistic constraints on the inference of word meaning in real word learners.

Key words: word learning, cross-situational learning, Quine's Problem

1. Word learning and indeterminacy of meaning

Children are prolific word learners, learning around 60,000 words by age 18 (Bloom, 2000). Their prodigious word-learning abilities are even more

---

remarkable when we consider some of the challenges facing the word learner, including the need to segment words from connected speech (Saffran et al., 1996), to generalise word forms across speakers (Henderson and Graham, 2005), and to identify the syntactic properties of those words (Mintz, 2002). In this paper we focus on another aspect of the word learning problem: inferring word meaning. Children will typically encounter words in a complex environment. How do they know what these words mean? Every time a word is used, there may be many meanings which a learner could infer as the word's true meaning: the learner will face *referential uncertainty*. As discussed below, a widespread observation in the literature is that there are potentially *infinitely* many candidate meanings which would be consistent with any given situation of usage. This idea, which we will refer to as *infinite referential uncertainty*, is commonly attributed to Quine (1960), although on our reading (discussed below) we think Quine's central point was rather different.

Regardless of its provenance, the infinite referential uncertainty hypothesis has been crucial in the development of two approaches to word learning, which differ in emphasis but are entirely compatible in content. One position emphasises the importance of heuristics which guide word learning, serving to reduce referential uncertainty and allow the learner to make accurate inferences about word meaning. These heuristics might include: exploiting the attentional focus of a speaker (Tomasello and Farrar, 1986); the assumption that words refer to whole objects (Macnamara, 1972); using knowledge of the meaning of other words to constrain hypotheses about the meaning of a new word, for example by assuming that words have mutually exclusive meanings (Markman and Wachtel, 1988); using argument structure and syntactic context to constrain the meaning of new words (Gillette et al., 1999). Heuristic-driven accounts emphasise how such constraints enable learners to eliminate uncertainty about word meaning and form good hypotheses about word meaning on even a single exposure to a word.

In the strongest accounts, these heuristics are hypothesised to eliminate all uncertainty. However, the possibility that word learners may be confronted with *some* residual referential uncertainty even after these heuristics have done their work has driven a recent burst of interest in a second approach to word learning, emphasising integration of information across multiple exposures as a means for learning in the face of referential uncertainty. *Cross-situational learning* comes in various flavours, from the classic formulation provided by e.g. Siskind (1996) to associationist treatments (Yu and Smith, 2007) to more minimal accounts (Smith et al., 2011; Medina

et al., 2011). For instance, in its most powerful instantiation (e.g. Siskind, 1996), cross-situational learning involves tracking the set of meanings which has been consistently inferred on every exposure to some target word: the word's true meaning should be a member of this set, which can be winnowed down across a series of exposures until it includes only the true meaning. Cross-situational learning accounts typically assume the presence of heuristics which serve to reduce referential uncertainty to manageable levels: rather than replacing heuristics, the contribution of this research is to explore the extent to which word learning is possible even given some residual (i.e. non-zero, but typically small) referential uncertainty.

Focussing on the interaction between heuristic and cross-situational approaches, in previous work (Blythe et al., 2010) we applied mathematical techniques to quantify what residual level of referential uncertainty a cross-situational learner can tolerate and still learn a large lexicon in a reasonable timeframe. Our previous work focussed on calculating learning times for lexicons given finite meaning spaces and finite levels of referential uncertainty. In this paper we apply similar techniques to tackle the problem of cross-situational learning for infinite meaning spaces under infinite referential uncertainty. In doing so, we seek to address what is often (perhaps rather loosely) called "Quine's Problem" or "the gavagai problem", the notion that word learning under infinite referential uncertainty is impossible. We show that word learning under such conditions is in principle possible, provided that learners have heuristics which at least rank the plausibility of each candidate meaning at every exposure. Thus, as in fact envisaged by Quine in his exposition of the indeterminacy of translation, word learning is possible if learners know, of the infinitely many possible meanings a word could have on any given situation of usage, how plausible each of those meanings are, and that some are more plausible than others. Within this very general set of conditions, given enough time, cross-situational learning can be used to eliminate uncertainty. Furthermore, cross-situational learning will in principle be possible even if the learner's heuristics only impose very weak constraints on the ranking in terms of plausibility. This work therefore suggests similar conclusions to our previous work exploring finite referential uncertainty: word learning heuristics can in principle be far weaker than previously suggested and still allow word learning — in fact, those heuristics can be so weak as to admit *infinitely many* possible meanings on any given exposure to a word, which renders single-exposure word learning impossible. Importantly, we therefore directly overturn the commonly-held assumption that word learn-

3

ing is impossible in the face of infinite referential uncertainty. Furthermore, the fact that word-learning heuristics which provide only weak constraints on possible word meanings can nonetheless allow word learning has potential implications for our understanding of the heuristics and cognitive biases underpinning word learning, and therefore on the empirical research attempting to uncover those biases. Firstly, this moves the explanatory burden from 'smart' inference by learners to 'dumb' crunching of cross-situational statistics, therefore requiring us to assume less of word learners in terms of their ability to accurately infer word meaning. Second, word learning heuristics do not need to allow learners to make good guesses on a single exposure to a word, which is a standard diagnostic in experimental research: weaker, unreliable, probabilistic heuristics can also play a key role, and therefore merit investigation.

2. 'Quine's Problem': learning under infinite uncertainty

Words are used in complex environments, and each word could label any part of that complex environment. Worse, words can label objects and events which are not perceivable to speaker or hearer (e.g. events which are spatially or temporally distant from the time of speaking). And this is only considering the obvious possibilities — words might have 'strange' meanings (e.g. featuring disjunctions of the meanings of 'normal' words, meaning for instance "a spark plug *or* an elephant", "happiness *or* the number 17", etc). This idea, commonly attributed to Quine's work on radical translation (of which more below), appeals to the notion that on any situation there will be *infinitely many* possible meanings that a novel word could have:

> "Even if we restrict ourselves to middle-sized objects ... we are stuck with Quine's problem, which is that children who hear a word and know that it refers to a rabbit are still faced with an indefinite number of possible meanings for this word" (Bloom, 2000, p. 56)

> "Quine (1960) points out that there are an infinite number of true facts about the world that a learner might need to entertain as potential meanings of each utterance." (Siskind, 1996, p. 45)

> "Worse, or so philosophers tell us, learners might conjure up absurd and endlessly differing representations for those entities we adults call 'the cats.'" (Gillette et al., 1999, p. 136)

"Famously articulated by Quine (1960), in any naming situation there are infinite interpretations for an unknown word. Thus, children face a daunting task of ambiguity resolution that they must solve thousands of times." (McMurray et al., 2012, p. 831)

"Word learning is often described as a difficult task because the world offers infants a seemingly infinite number of word-to-world mappings in just one moment in time (Quine, 1960)." (Vlach and Johnson, 2013, p. 375)

"Determining the meaning of a newly encountered word should be extremely hard, due to the (in principle, unlimited) referential uncertainty inherent in the task (Quine, 1960)." (Smith et al., 2011, p. 480)

Such claims about infinite referential uncertainty are widespread in the literature, and have played an important role in the development of the theoretical motivation for research on the heuristics children use to eliminate uncertainty during word learning: seminal papers on the Mutual Exclusivity constraint (Markman and Wachtel, 1988), the shape bias (Landau et al., 1988), joint attention (Baldwin, 1991), or lexical constraints in general (Golinkoff et al., 1994) make explicit reference to the problem of there being "infinitely many" or "limitless" meanings a word could have. The consensus is that word learning is impossible given this infinite uncertainty, and that heuristics are required to eliminate some of these candidate meanings. In this paper we explore the validity of this widely-held and entirely reasonable intuition, and in particular show that it does not hold in a wide range of well-defined circumstances. However, before doing so it is worth briefly considering whether Quine's Problem was actually posed by Quine.

3. Quine on word learning

Quine (1960) introduces the problem not in terms of word learning, but in terms of "radical translation", his examination of how the language of a "hitherto untouched people" can be translated. Much of his discussion focusses on issues which are only tangentially relevant to word learning (such as how we can understand signs for agreement or disagreement from our informant) or which are relevant to language learning in general but not to referential uncertainty (e.g. the order in which types of meaning must be

learnt). The key passage, referred to implicitly in much of the literature, is as follows:

> "For, consider gavagai. Who knows but what the objects to which this term applies are not rabbits after all, but mere stages, or brief temporal segments, of rabbits? In either event the stimulus situations that prompt assent to Gavagai would be the same as for Rabbit. Or perhaps the objects to which gavagai applies are all and sundry undetached parts of rabbits; again the stimulus meaning would register no difference." (Quine, 1960, p. 51-52)

Interestingly, Quine never explicitly states that there are an infinite number of possible meanings for each word at any episode, although this is an entirely reasonable inference from his discussion of what we referred to earlier as 'strange' meanings. Rather, it seems to us that Quine is much more concerned with the fact that some candidate meanings are *in principle always indistinguishable* from a word's true meaning because they are always applicable to the same stimuli: "Point to a rabbit and you have pointed to a stage of a rabbit, to an integral part of a rabbit, to the rabbit fusion, and to where rabbithood is manifested" (Quine, 1960, p. 52). As we note below, this renders word learning impossible in principle, if we treat word learning as a process of eliminating spurious meanings, as theories of cross-situational learning typically do; in other words, we accept Quine's central point that, if two candidate word meanings are always indistinguishable, cross-situational learning cannot distinguish between them. To escape this conundrum, we have to assume that some of these indistinguishable meanings (i.e. "rabbit") are just more plausible than others ("undetached rabbit parts"). Quine suggests that we can't assume this *a priori*, although he acknowledges that in practice we do rank candidate meanings in terms of plausibility, saying that the linguist assumes that "the native is enough like us to have a brief general term for rabbits ... [and not] for rabbit stages or parts" (p. 52). In the remainder of this paper, we show that Quine's intuitions on plausibility ranking over candidate word meanings solves the problem of word learning in the face of infinite referential uncertainty, and quantify how weak those rankings can be.

4. Model of the learning problem

Following the approach taken by other modelling and experimental work on cross-situational learning, we characterise the word learning process as one of eliminating uncertainty about word meaning, and we seek to identify conditions under which the meaning of a word can be uniquely determined after multiple exposures. We take an ideal observer approach (Frank and Tenenbaum, 2011) and consider an optimal cross-situational word learner attempting to identify the meaning of a single word; we measure time t in terms of the number of exposures to this word that the ideal learner has received. Each time the word is uttered, the learner infers (by applying their battery of word learning heuristics to the situation in which the word is used) one or more possible discrete meanings for that word drawn from an infinite set of meanings according to some specified probability distribution. We make two assumptions regarding the form of this distribution:

Assumption 1: The word's true meaning, the *target meaning*, is always included as one of the inferred meanings.

Assumption 2: All other meanings (the *incidental meanings*) have some nonzero probability of *not* being inferred.

Assumption 1 is often identified as a problem in theories of cross-situational learning (see e.g. Gleitman, 1990) — although it simplifies our analysis, we show in Section 6 that word learning is possible if we violate Assumption 1 and allow that the word's true meaning is not always (or even not often) inferred.

Assumption 2 corresponds to the requirement that not all possible incidental meanings are inferred on every exposure to a word, which would clearly render cross-situational learning impossible. Furthermore, it corresponds to (our reading of) Quine's concern, discussed above, that some meanings are in principle indistinguishable from the word's true meaning: were this the case, Assumption 2 would be violated for those meanings. We consider this assumption to be necessary for cross-situational word learning to be possible, and probably necessary for any theory of word learning which conceives of learning as a process of converging on a correct meaning at the expense of all alternatives.

These assumptions about the candidate meanings a learner infers on a single exposure to a word therefore constitute our model of the learner's word

learning heuristics: manipulating the probability distribution over inferred meanings allows us to simulate learners with very powerful word learning heuristics (e.g. when the learner always infers the target meaning and a very small number of incidental meanings) or far weaker heuristics (e.g. where the number of incidental meanings inferred on any episode is large or even infinite, due to the learner's inability to make strong inferences about the word's true meaning). Note that, unless we assume that the learner's heuristics are maximally strong and eliminate *all* referential uncertainty (i.e. no incidental meanings are inferred, an extreme version of our Assumption 2), cross-situational learning still has to do some work: every exposure will feature multiple candidate meanings, and the target meaning will therefore not be identifiable on any single exposure.

Our model is agnostic about the source of these constraints on possible word meanings, but we envisage them as arising through the interplay between the learner's word learning heuristics and the (linguistic, social and physical) context in which a word is used: for instance, on hearing a word, a learner may assume that it is likely to be a noun given the syntactic context in which it occurred (Gillette et al., 1999), that it is likely to refer to whole objects rather than part of an object (Macnamara, 1972), that it is likely to refer to one of the cluster of objects that the speaker was attending to (Tomasello and Farrar, 1986), and that, given all the foregoing, it is likely to refer to the toy that just beeped and flashed in a rather salient manner.

Other than these two assumptions given above, we do not place any restrictions on the meanings that are inferred or the relationships between the inferences that are made within or between exposures. It should be noted that assuming discrete meanings, as we do here, does not commit us to any assumptions about the nature of meanings or the relationships between meanings. For instance, meanings could be interpreted as existing in a hierarchically- and similarity-structured space, which would be captured in our model by manipulating the set of incidental meanings associated with the word's true meaning and the distribution from which those incidental meanings are drawn, such that incidental meanings which are related to the target are more likely to be inferred by the learner, and more general meanings are more likely to be inferred than more specific, restricted meanings. We return to word learning for such structured meaning spaces in the discussion. Relatedly, in Sections 5.3 and 5.4 we explore scenarios where there are correlations between meanings, either within a single exposure (the fact that one meaning has been being inferred on a given situation could increase,

8

or decrease, the probability that other meanings are inferred, for example if those meanings are related) or across exposures (the inference of a meaning in one exposure could increase or decrease the probability of its being inferred in subsequent exposures, for example as a result of the temporal structure of dialogue).

We assume that our idealised learner uses the full cross-situational learning strategy defined in e.g. Siskind (1996) or Blythe et al. (2010), whereby the learner keeps track of the set of meanings that have been inferred in *every* exposure up to time t. Due to Assumption 1, the target meaning is always part of this set; if the sequence of exposures is such that this set comprises precisely one meaning (which must be the target meaning), the word is taken to be learnt. In Section 6, where we relax Assumption 1, we necessarily adopt a less restrictive, and consequently less powerful, conception of cross-situational learning.

Our model of word learning is sketched in Fig. 1. Given this very simple model, what are the formal conditions under which word learning is possible in finite time? Given the standard interpretation of Quine discussed above, we are particularly interested in scenarios where words can be learned in finite time despite potentially infinite referential uncertainty. We place no restrictions on how small learning times must be: in previous work (Blythe et al., 2010) we investigated learning times for large lexicons in timescales which were comparable to those likely to be available for real language learners, but since here we are concerned with exploring the feasibility of cross-situational learning *in principle*, learning in finite time is our only requirement: learning in practicable timescales, as we explore for the finite meaning case in Blythe et al. (2010), is an additional constraint and a worthwhile avenue for future work.

5. Formal conditions for word learning in finite time

5.1. A simple formulation of cross-situational learning

We define the learning time $t^*(\epsilon)$ that is needed for the meaning of a word to be uniquely identified with probability $1 - \epsilon$, where ϵ is some small constant. A probabilistic notion of learning time is necessary because it is always possible to construct an arbitrarily long sequence of exposures where the target word remains unlearned (e.g. any sequence where the target meaning and one particular incidental meaning are both inferred: such sequences constitute temporary instances of the problematic case that concerned Quine,

9

Figure 1: A sketch of our model of word learning. For every word (here, *horse*) there is a target meaning (represented here by the silhouette of a horse) and an infinitely-large set of possible incidental meanings (represented here by silhouettes of other animals). Learning consists of a series of episodes; on each episode the learner samples meanings from the set consisting of the target meaning plus the incidental meaning, with these meanings being sampled with differing probabilities (e.g. if our Assumptions 1 and 2 are followed, the target meaning is always sampled, and every incidental meaning has some probability of *not* being sampled at any given episode). This sampling procedure represents the process whereby the learner infers likely word meanings based on their word learning heuristics and the context in which the target word is used. On the basis of each such learning episode, the learner updates their set of candidate meanings for the target word: on the first exposure this will simply be the set of inferred meanings, each of which is equally likely to be the word's true meaning; over episodes this set of candidate meanings is winnowed down as candidates fail to occur in the set of inferred meanings and can therefore be eliminated. Eventually, if cross-situational learning is possible, the set of candidate meanings will consist of a single meaning, the target meaning, and the word has been learnt.

namely where two candidate word meanings cannot be differentiated). The learning time is finite when the probability of such vexatious sequences is sufficiently small that $t^*(\epsilon) < \infty$ for any arbitrarily small, but nonzero, ϵ.

To understand this criterion in more detail, it is helpful to think of a population of learners, each drawing their own set of inferences of a word's meaning from a common distribution. Then, ϵ can be interpreted as the fraction of all learners who would not have learnt the target word's meaning at time $t = t^*(\epsilon)$. Clearly, if $t^*(\epsilon_0)$ becomes infinite at some finite ϵ_0, then some fraction of learners fail to learn the word in a finite amount of time. Furthermore, once the population size N exceeds $1/\epsilon_0$, then there is typically at least one agent in the population who fails to learn the word. Our criterion thus amounts to insisting that *all* agents learn the word in a finite amount of time (although this can be very long), and we consider it preferable to using other statistics, such as the mean time to learn the word (which can be infinite, even though every member of the population can learn the word in a finite amount of time) or the median (which can be finite even when nearly half the population fail to learn the word in a finite amount of time).

As outlined above, we assume that meanings are discrete, and that there are infinitely many possible incidental meanings (i.e. there are infinitely many meanings which *could*, in principle, be inferred on any occurrence of the target word). We therefore have a one-to-one mapping between positive integers m and incidental meanings. We define $\mathcal{K}(t)$ as the set of incidental meanings that have been inferred in *every* episode up to time t; the word is learnt when $\mathcal{K}(t^*) = \emptyset$, i.e. when all incidental meanings have been eliminated, and the word's true meaning is therefore revealed. The crucial quantity in determining whether the learning time is finite or not is the *residual plausibility* of an incidental meaning m, which is defined as follows

$$p_m(t) = \Pr\left[m \in \mathcal{K}(t) | \mathcal{F}_{m-1} \cap \mathcal{K}(t) = \emptyset\right] \tag{1}$$

where $\mathcal{F}_m$ is the set of meanings $\{1, 2, \ldots, m\}$. In words: the residual plausibility of incidental meaning m at time t is the probability that it is still a candidate meaning for the word *given that* the first $m-1$ incidental meanings have all been excluded after t exposures. This definition can be applied for any ordering of the meanings; however it is often convenient to assume that they are arranged in decreasing order of their probability of being inferred in any given episode.

Given this definition, we can now state our main results, a formal derivation of which is given in the Appendix. First, the probability $L(t)$ that the

11

word is learnt by time t is simply the probability that all incidental meanings have been eliminated, i.e. $\mathcal{K} = \emptyset$. We find that this can be written in terms of the residual plausibilities as

$$L(t) = \prod_{m=1}^{\infty} [1 - p_m(t)] \tag{2}$$

i.e. the probability of learning a word is given by the probability that the first non-target meaning has been excluded, multiplied by the probability that the second non-target meaning has been excluded conditioned on the first being excluded, multiplied by the probability that the third non-target meaning has been excluded conditioned on the first two being excluded, and so on down the infinite hierarchy of non-target meanings.

Then, we find that the learning time $t(\epsilon)$ is finite if and only if

$$\lim_{t \to \infty} \left[\sum_{m=1}^{\infty} p_m(t) \right] = 0 \tag{3}$$

i.e. learning time is only finite if the residual plausibility of *all* incidental meanings reaches 0.

Informally, we can interpret this result as follows: if the combined residual plausibility of all incidental meanings vanishes in the limit of an infinite sequence of exposures, the meaning of a word can be identified with arbitrarily high probability in a finite time. Blythe et al. (2010) considered cases where the set of possible incidental meanings was finite: as is made clear in the formulation above, the learning time in such cases will be finite as long as there is some probability that each incidental meaning is excluded at any given exposure, which follows from our Assumption 2. Therefore, under our assumptions, word learning is always possible when the set of possible incidental meanings is finite. The case of an infinite set of incidental meanings is best elucidated by means of explicit examples, to which we now turn.

5.2. Learning times given infinite meaning spaces

Let us start by assuming that each incidental meaning m is inferred with a constant probability a_m in each exposure, and these inferences are statistically independent (that is, the probability that both meanings m and m' are inferred in any given exposure is just $a_m a_{m'}$, and is the same in each episode). Under such circumstances, the probability that meaning m is inferred in all

12

of the first t episodes is a_m^t, and this does not depend on whether any other meanings (e.g. meanings $1, 2, \ldots, m-1$) have been inferred in each of those t exposures or not. In this case, the residual plausibility of meaning m at time t is $p_m(t) = a_m^t$. Now, since by our Assumption 2 all $a_m < 1$, i.e. all incidental meanings have some non-zero probability of not being inferred on any given exposure, it follows that $p_m(t) \to 0$ as $t \to \infty$: that is, every incidental meaning becomes ever more implausible as time goes on. However, unlike in the case of finite sets of incidental meanings, this does not in itself guarantee a finite learning time, as we shall now see.

Let us arrange the meanings in decreasing order of their inference probability, i.e., such that $a_{m+1} \leq a_m$. Suppose that, for all incidental meanings, their probability of being inferred on any given exposure follow a power law: $a_m = cm^{-\gamma}$, where c and γ are positive constants. c denotes the probability of inferring the most frequent incidental meaning on any given episode, and must be less than 1 by our Assumption 2. γ denotes the rate of decay of the power-law distribution: for high γ, a_m decreases rapidly with m (i.e. only the few most frequent meanings have any substantial probability of being inferred on any given trial); as γ is reduced, the distribution of a_m flattens out, such that even lower-ranked meanings have a substantial probability of being inferred in any episode. Under these conditions, the sum in (3) converges (i.e. approaches some constant) for any $t > 1/\gamma$ and furthermore vanishes (i.e. converges to 0) in the limit $t \to \infty$. Importantly, note that the rate of decay of the inference probabilities can be very slow, for example, if γ is very close to zero. Indeed, when $\gamma \leq 1$, the mean number of meanings inferred in each episode is *infinite*: the infinitely-long tail of meanings, each with substantial probability of being inferred at each instance of use of the target word, leads to a situation where the learner is confronted with infinite referential uncertainty; nevertheless, the power-law decay of a_m is still sufficiently fast for a cross-situational learner to be able to eliminate all non-target meanings in finite time. This refutes the common intuition, discussed above, that cross-situational learning cannot work in principle in the face of infinite referential uncertainty.

This can be illustrated using simulations of cross-situational learning. Since it is impossible to use an infinite meaning space in a simulation, our approach here is to vary the number of available meanings, M, and investigate what happens as M increases (using $c = 0.99$ and $\gamma = 0.01$). In Figure 2, we plot the time at which all but a fraction ϵ of learners have learnt the word: this time is independent of M for sufficiently small ϵ.
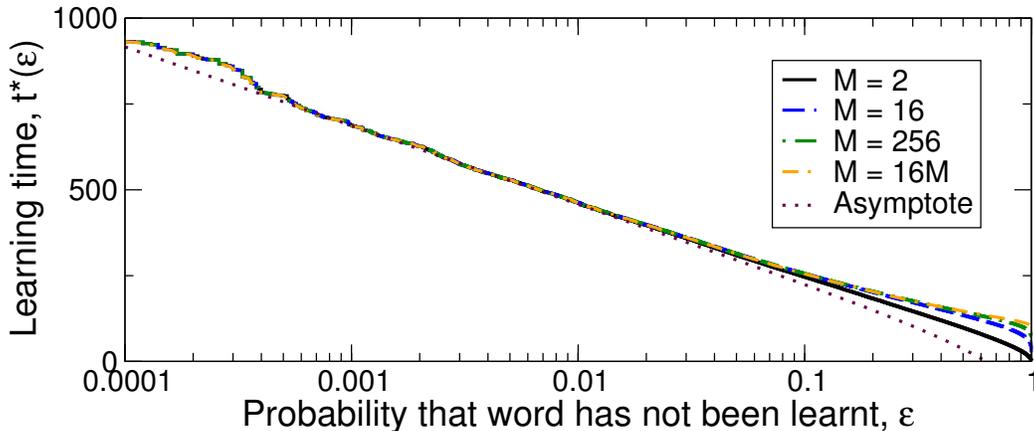
Figure 2: Time to learn a word with probability $1 - \epsilon$ with confounder frequencies governed by the power-law $a_m = cm^{-\gamma}$, with $c = 0.99$ and $\gamma = 0.01$. Note that, crucially, learning times are independent of the number of incidental meanings M, as indicated by the invariance of learning time as we vary M. For $\epsilon < 0.01$, the learning times are well approximated by the asymptotic result $t^*(\epsilon) \sim \ln(\epsilon)/\ln(c)$. The values of M used in the simulation are powers of 2: $2^1 = 2, 2^4 = 16, 2^8 = 256$ and $2^{24} = 16,777,216$ (denoted 16M in the figure).

Note that for the exponent $\gamma = 0.01$ used in the plot, the mean number of meanings inferred on any given episode is $M^{1-\gamma}$, i.e., $M^{0.99}$, meaning that virtually all meanings that could be inferred are inferred on every instance. Despite this, if we are only interested in the most laggardly learners, only the most frequent of these many incidental meanings is relevant to learning time: by the time the most problematic incidental meaning has been eliminated by the slowest learner, the other less frequent incidental meanings will also have been eliminated. The influence of the most frequent confounder on learning times is a feature of all the cases we consider in this paper. In fact, in the regime $\epsilon < 0.01$, we find good agreement between our simulation results and a very crude approximation to (2) in which we consider only the most frequent confounder, which is inferred on every episode with probability c, i.e.,

$$L(t) \approx 1 - a_1^t = 1 - c^t . \qquad (4)$$

Setting this equal to $1 - \epsilon$ we obtain an estimate for the learning time $t^*(\epsilon)$. Here, we find that that, as $\epsilon \to 0$, $t^* \sim \ln(\epsilon)/\ln(c)$. As can be seen from Figure 2, the agreement with the numerical data for $\epsilon < 0.01$ is excellent.

14

Given that we find that a word can be learnt even when the plausibility ranking of non-target meanings is very weak, one may legitimately ask whether it is sufficient only for a_m that decreases to zero as $m \to \infty$ to guarantee a finite learning time. It turns out that this is not the case. In the Appendix we provide the example of a set of frequencies that decay logarithmically, and find that the word is not necessarily learnt in finite time. In this case, whether learning times remain small or diverge towards infinity depends on the size of the set of incidental meanings: for logarithmic decay, it turns out that learning times are determined by the most frequent confounder, as above, when this set is less than 10^{18}, a vast set of possible incidental meanings. Only above this size may learning times that diverge as M is increased become apparent. This example shows that although a cross-situational learner can in principle fail to learn the meaning of a word in a finite time when the inference probabilities decay logarithmically, learning is nevertheless possible unless the number of meanings that may be inferred is extremely large.

5.3. Correlated meaning inferences within episodes

It is instructive to consider cases where the inference of certain meanings enhances (or diminishes) the probability that certain other meanings are inferred: in our simple model of meaning, such correlations between meanings can be used to capture hierarchical- or similarity-based structure within the meaning space; as such, cross-situational learning in the presence of such correlations is an important topic of study. An extreme (but illustrative) case is where the inference of incidental meaning m on any given episode implies that all meanings that are more likely to be inferred are also inferred. If we order the meanings according to decreasing probability of being inferred, this means that whenever meaning m is inferred, all meanings $m' < m$ are also inferred. Conversely, if meaning m is not inferred, all meanings $m' > m$ are also not inferred (see next paragraph for the relationship between such meanings and Quine's 'strange' meanings).

Under this scenario, the residual plausibility of all meanings $m > 1$ (i.e. other than the most probable incidental meaning) is zero. This is because if all meanings $m' < m$ have been excluded, then meaning m must also have been excluded. Hence, the probability the word is learnt after time t is just equal to the probability that the most probable incidental meaning has been excluded. All that is needed for a finite learning time is for this probability to approach unity as $t \to \infty$. These within-episode correlations therefore lead

to *faster* learning than when meanings within a single episode are statistically independent. More generally, any incidental meaning m whose inference implies the inference of a (necessarily more plausible) meaning has residual plausibility $p_m(t) = 0$, and can therefore be ignored for the purposes of computing learning times. All such *shadowed* meanings are therefore invisible to a cross-situational learner. Quine's 'strange' meanings seem excellent candidates for shadowed meanings: if "undetached rabbit parts" is inferred, then "rabbit" would necessarily also be inferred and would be more plausible.

What about weaker within-episode correlations, which do not produce shadowed meanings? Learning times here should be somewhat similar to the case where each incidental meaning is statistically independent of any other, since weakening of within-episode correlations gradually approaches this statistically independent case. While we do not have a general formation for these weak-correlation scenarios, we can investigate numerically. There are of course a large number of ways that we could construct such meaning spaces. We use a simple Markov chain model, in which meaning 1 is inferred with probability c, and then meaning $m > 1$ is inferred with probability p_m if meaning $m - 1$ is inferred, or with probability q_m if meaning $m - 1$ is not inferred. Specifically, we choose

$$p_m = \frac{(1 - \alpha)(m - 1)^\gamma + \alpha c}{m^\gamma} \quad \text{and} \quad q_m = \frac{\alpha c}{m^\gamma} \tag{5}$$

so that the probability that any particular meaning m is inferred is $cm^{-\gamma}$, as previously. However, given that a meaning m is inferred, there is an increased likelihood that meanings with a similar plausibility are inferred, depending on α. Specifically, if $\alpha = 1$, $p_m = q_m$, which means that each meaning is inferred independently of any other meaning, which is the special case we treated before. On the other hand, as $\alpha \to 0$, the probability $p_m \to 1$ as $m \to \infty$, which means that long strings of infrequent meanings appear together in this limit.

As anticipated, we find that the learning time $t^*(\epsilon)$ is largely unaffected by the presence of correlations, even for small values of α, as shown in Figure 3. In particular, we find the same approximation that captured the behaviour of the statistically-independent case, $t^*(\epsilon) \sim \ln(\epsilon)/\ln(c)$ as $\ln(\epsilon) \to 0$, also works well here: even when within-episode correlations are present, the slowest learners are those who retain the most plausible confounding meaning as an alternative hypothesis for an extended time.
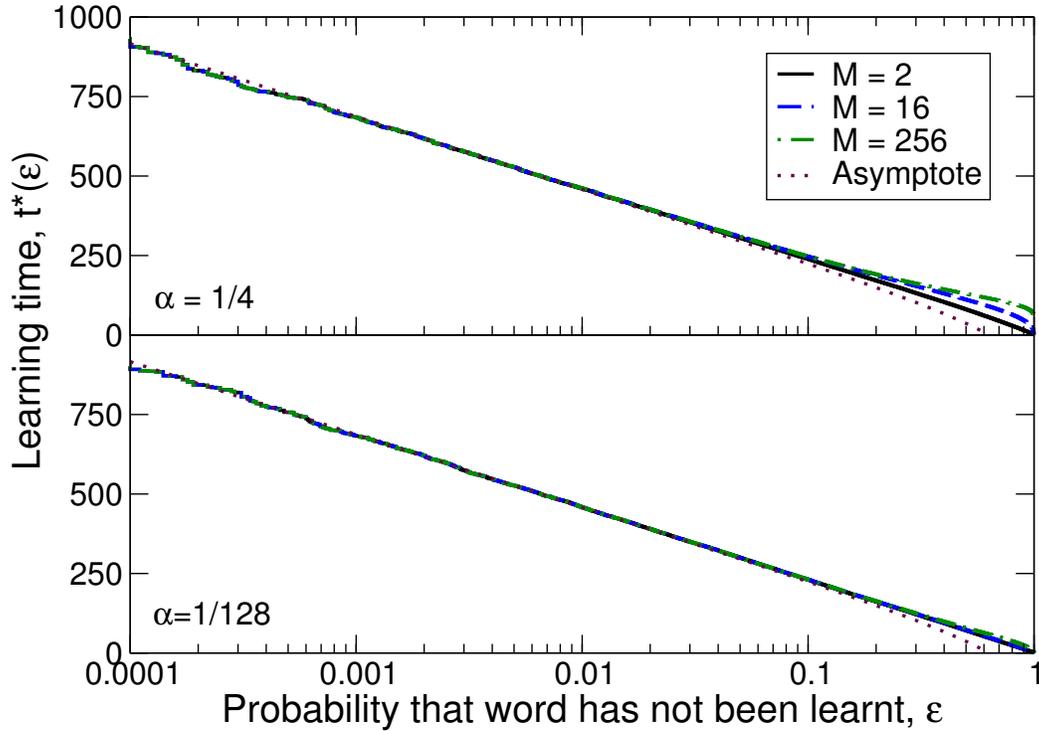
16

Figure 3: Time to learn a word with probability $1 - \epsilon$ with confounder frequencies governed by the power-law with $c = 0.99$ and $\gamma = 0.01$. Meanings are correlated so that strings of neighbouring meanings are likely to appear, the degree of correlation being determined by α. As in Fig. 2, learning time is independent of the number of incidental meanings, M. Furthermore, learning time is largely unaffected by correlations between meanings, as can be seen by comparing the upper and lower panels of this figure (within-episode correlations are substantially stronger in the lower panel), and indeed by comparing this figure to Fig. 2 (where there are no within-episode correlations): any differences from the uncorrelated case (Fig. 2) are confined to the large-ϵ regime.

5.4. Correlated meaning inferences across episodes

Inferences mighty also be correlated in time: a meaning being inferred in one episode may enhance (or diminish) its probability of being inferred in the next episode, which would seem likely given that words are used in dialogues where certain themes and linguistic constructions recur (Pickering and Garrod, 2004). With such correlations, one can typically associate a timescale τ after which the correlation has decayed. In this case, we would simply expect all learning timescales to be increased by a factor of order τ. This does not affect whether the learning time is infinite or not, unless τ itself is infinite.

6. Failure to infer the target meaning: relaxing Assumption 1

In the foregoing models, we assumed that the word's true meaning, the target meaning, was inferred at every episode (our Assumption 1). While this significantly simplifies the formal analysis of cross-situational learning, it turns out (contra e.g. Gleitman, 1990) not to be necessary for word learning to be possible. Rather, what is required is that the target meaning is the most likely to be inferred whenever the word is uttered: given enough exposures, this meaning will have been inferred more times than any arbitrarily high-probability incidental meaning, and a learner who identifies the meaning which is most strongly associated with the target word will learn the word (as discussed in, e.g., Smith et al., 2011). Therefore, cross-situational word learning is possible as long as learners employ heuristics that reliably lead to the target meaning being the most plausible (over many episodes, not necessarily in any given episode).

In order to show that this is the case, we relax our Assumption 1: rather than always being inferred, we assume that the target meaning is inferred with probability c', and the probability of the incidental meanings decays according to a power law as before, with the probability of the mth incidental meaning being inferred on any episode being $a_m = c'm^{-\gamma}$, m ranging from 2 (for the most plausible incidental meaning) to infinity. $c' = 1$ corresponds to our standard model, where Assumption 1 is met and the target meaning is always inferred; $c' < 1$ corresponds to cases where Assumption 1 is violated and the target meaning is not always inferred, including cases (for $c' < 0.5$) where it is not inferred on the majority of exposures. Due to the power-law decay, regardless of the value of c', all incidental meanings are less likely to be inferred than the target meaning.

In order for the word's true meaning to be identifiable despite the possibility of the target meaning being absent on one or more exposures, we have to relax our model of cross-situational learning somewhat. Under these circumstances, a strictly eliminative cross-situational learner would be forced to conclude that the target word had *no* meaning, since for every meaning (including the target meaning) there will be episodes where that meaning is not inferred. Instead, we assume a less strict form of cross-situational learning. Learners maintain a single candidate hypothesis about the word's meaning, which they revise after each exposure. If the candidate meaning is included in the set of meanings selected at exposure i, it is retained; otherwise, the single most frequent meaning included in the current exposure is selected as the new candidate (i.e. the meaning which has co-occurred most frequently, but not necessarily always, with the target word); in the event that there are multiple equally-frequent meanings, one is selected at random. In Blythe et al. (2010) we term a related form of cross-situational learning, where new candidate meanings are selected proportionally to their frequency of occurrence, *Approximate* cross-situational learning, to distinguish it from the maximally-strict eliminative form of cross-situational learning, and also from other forms of cross-situational which make less use of cross-situational statistics (e.g. the Minimal strategy from Blythe et al. 2010; Propose but Verify from Trueswell et al. 2013).

As can be seen in Fig. 4, the target word can still be learnt even if the target meaning is not reliably inferred (e.g. $c' < 1$) or even if it is not usually inferred ($c' < 0.5$) — the less reliably the target is inferred, the slower learning becomes, but since the target meaning is still the single most probable meaning, eventually the cross-situational statistics will allow the learner to identify the correct meaning. In other words, cross-situational learning will be possible as long as the learner's heuristics ensure that the target meaning is the most likely to be inferred *in the long run*, even when the target meaning is never unambiguously provided and even when the target meaning is often not inferred.

7. Discussion

We make two central assumptions in our initial model: that the target meaning is *always* inferred (our Assumption 1), and that all incidental meanings have some non-zero probability of *not* being inferred, i.e. unlike the target meaning, they are not always inferred (our Assumption 2). As we
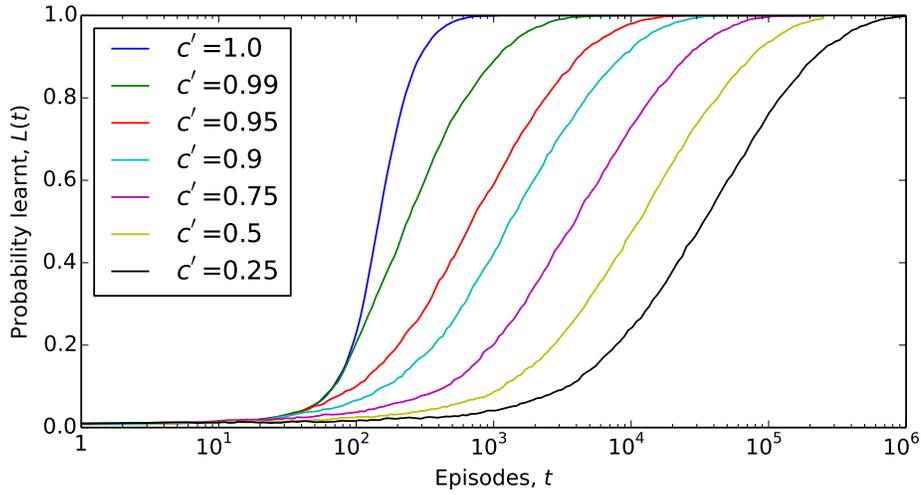
Figure 4: Proportion of 10,000 simulation runs in which the target word is learnt after a given number of exposures, with meaning frequencies governed by the power-law with various values of c', denoting the probability that the target meaning will be inferred on any given episode: $c' = 1$ corresponds to the case where the target is always inferred (satisfying our Assumption 1); learning is slower but still possible for $c' < 1$, corresponding to cases where the target meaning is not inferred on every trial, including for cases where $c' < 0.5$, i.e. where the target meaning is not inferred on the majority of trials. These results are for M=100; results for larger M are qualitatively similar.

show in Section 6, Assumption 1 is not required for cross-situational word learning to be possible: words can be learnt despite potentially infinite referential uncertainty even if a word's true meaning is not always, or not often, inferred. Our second assumption embodies Quine's observation, which we take to be necessary for *any* model of word learning, that incidental meanings which are in principle indistinguishable from the target meaning will block learning.

We have focussed here on the learning of a single word, but the same techniques can be extended to the learning of large lexicons, using the techniques presented in our previous work (Blythe et al., 2010; Reisenauer et al., 2013): in particular, Reisenauer et al. (2013) consider the case where words are not learnt independently, such that learning one word can facilitate the learning of another, through the Mutual Exclusivity constraint.

Our idealised learner in our initial model employs the most powerful form of cross-situational learning possible, tracking the set of meanings which have consistently co-occurred with the target word. In Section 6, we consider a weaker form of cross-situational learning, where learners track not the ever-present set of word meanings, but simply maintain some estimate of the frequency of candidate word meanings. Still weaker versions of cross-situational learning are possible, and indeed seem to better characterise human cross-situational learning in some scenarios (Medina et al., 2011; Trueswell et al., 2013): for instance, learners might simply retain a single preferred candidate meaning across exposures, persisting with or rejecting this hypothesis in the light of each exposure. We have previously shown, for the case of finite sets of incidental meanings, that weaker strategies tend to increase the time required to learn a lexicon, without introducing any qualitative shift in the conditions under which cross-situational learning is possible (Blythe et al., 2010). While exploring these mechanisms for the case of infinite meaning spaces is an area for future work, we expect that the picture presented here will in general hold: in particular, details of the exact cross-situational procedure applied are unlikely to influence whether learning time is finite or not.

Our treatment of meaning is rather minimal: meanings are simply discrete values drawn from a potentially infinite set. This extremely general treatment allows us to abstract away from any particular theory of word meaning, and as we mention when introducing the model, we expect that this same approach could be implemented in such a way that meanings could naturally be interpreted as being drawn from a hierarchically- and similarity-structured space, by manipulating the set of incidental meanings associated

with the word's true meaning and the distribution from which those inciden-
tal meanings are drawn, such that incidental meanings which are related to
the target are more likely to be inferred by the learner, and more general
meanings are overall more likely to be inferred than more specific, restricted
meanings. Note that, in the latter case, the cross-situational learner still
requires it to be the case that the word's true meaning is the most likely to
be inferred (see next paragraph) — biases such as a size principle (Tenen-
baum, 1999) or Mutual Exclusivity, which we take to be part of the learner's
heuristics, would be required to ensure that words with specific, restricted
meanings could be successfully learnt. Verifying that this is the case would
be a valuable addition to the initial model we outline here.

All of the models outlined above assume that a word's true meaning is the
most likely meaning to be inferred on any one exposure (although it may not
actually be inferred), and therefore, averaged over many exposures, will be
the most frequently inferred meaning for that word, allowing cross-situational
learning to take place. This assumption bears some discussion. Firstly, we
have assumed in our models that the probability distribution over the target
and incidental meanings remains constant over the course of learning, but this
is not necessarily the case. For instance, for some word meanings which rely
on learning *a priori* more probable meanings first and then applying Mutual
Exclusivity, the target meaning may only be learnable once the learning of
other words re-orders the plausibility of candidate meanings such that the
target meaning becomes the most plausible remaining meaning. In other
words, we expect that word learning will still be possible if the target word
is *eventually* the most plausible. Secondly, the fact that words are socially
learned and culturally transmitted provides some guarantee that the types
of meanings learners tend to infer will indeed be the types of meanings that
words come to have: a range of computational and experimental work shows
that languages adapt to the biases of language learners as a result of their
repeated learning and transmission (e.g. Kirby, 2000; Griffiths and Kalish,
2007; Kirby et al., 2008): while the structure of the environment also feeds in
to this process (Perfors and Navarro, 2014), we should therefore expect that
word meanings will become adapted to the heuristics that learners use to
infer them; in the limiting case, a word whose meaning could not be inferred
by learners would be guaranteed to change in meaning. While the process
by which languages become adapted to the biases of language learners has
been well studied in the general case, explicit models of the process for word
learning would be worthwhile.

Finally, to reiterate: although the results we present results require that, in the long run, the word's true meaning is the most likely to be inferred, there is still substantial work for cross-situational learning to do. Firstly, although the learner is more likely to infer the word's true meaning than any other, they still face substantial or indeed infinite referential uncertainty: on any given exposure, the target meaning is merely one of an infinite number of inferred meanings, and cross-situational learning is required to eliminate this uncertainty. Secondly, as we show in Section 6, the target meaning does not have to be reliably inferred to be learnt. Cross-situational learning therefore turns an apparently impossible learning problem — inferring the meaning of a word from a series of exposures where each exposure offers infinitely many candidate meanings, possibly not including the true meaning — into a tractable one.

8. Conclusions

We have established a very general formal foundation for the study of cross-situational learning, obtaining an exact expression for the probability, $L(t)$ that a word is learnt after t exposures by an ideal cross-situational learner, valid for an arbitrary distribution of incidental meanings. We have also established criteria for when this learning time is finite, i.e. cross-situational learning is possible, based on the residual plausibility of an incidental meaning. When confounding meanings are inferred independently, the learning time is well estimated by the asymptotic formula $t^* = \ln \epsilon / \ln c$, where c is the frequency of the most common incidental meaning. This approximation also works well when there are within-episode correlations between inferred meanings. Importantly for the debate in the cognitive sciences over word learning in the face of infinite uncertainty, we have identified at least one scenario in which referential uncertainty at every exposure is infinite, yet learning times are still finite. This finding suggests that the common intuition, that cross-situational learning is impossible in such circumstances, is incorrect. Furthermore, we show that cross-situational learning is possible even if the learner's heuristics only impose very weak constraints on the plausibility ranking of possible meanings and even when the word's true meaning is not always (or even not often) inferred. This work therefore suggests that word learning heuristics can in principle be far weaker than previously suggested and still allow word learning, and therefore suggests that weaker, unreliable, probabilistic heuristics can play an important role; exploring such

biases in real word learners is therefore a worthwhile empirical aim.

References

Baldwin, D. A., 1991. Infants' contribution to the achievement of joint reference. Child Development 62 (5), 875–890.

Bloom, P., 2000. How Children Learn the Meanings of Words. MIT Press, Cambridge.

Blythe, R. A., Smith, K., Smith, A. D. M., 2010. Learning times for large lexicons through cross-situational learning. Cognitive Science 34, 620.

Frank, M. C., Tenenbaum, J. B., 2011. Three ideal observer models for rule learning in simple languages. Cognition 120, 360–371.

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.

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.

Griffiths, T. L., Kalish, M. L., 2007. Language evolution by iterated learning with Bayesian agents. Cognitive Science 31, 441–480.

Henderson, A. M. E., Graham, S. A., 2005. Two-year-olds' appreciation of the shared nature of novel object labels. Journal of Cognition and Development 6, 381–402.

Kirby, S., 2000. Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In: Knight, C., Studdert-Kennedy, M., Hurford, J. (Eds.), The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form. Cambridge University Press, Cambridge, pp. 303–323.

Kirby, S., Cornish, H., Smith, K., 2008. Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences, USA 105 (31), 10681–10686.

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.

McMurray, B., Horst, J. S., Samuelson, L. K., 2012. Word learning emerges from the interaction of online referent selection and slow associative learning. Psychological Review 119, 831–877.

Medina, T. N., Snedeker, J., Trueswell, J. C., Gleitman, L. R., 2011. How words can and cannot be learned by observation. Proceedings of the National Academy of Sciences, USA 108, 9014–9019.

Mintz, T. H., 2002. Category induction from distributional cues in an artificial language. Memory and Cognition 30, 678–686.

Perfors, A., Navarro, D. J., 2014. Language evolution can be shaped by the structure of the world. Cognitive Science 38, 775–793.

Pickering, M. J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27, 169–225.

Quine, W. V. O., 1960. Word and Object. MIT Press, Cambridge.

Reisenauer, R., Smith, K., Blythe, R. A., 2013. Statistical mechanics of lexicon learning in an uncertain world. Physical Review Letters 110, 258701.

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

Scott, D. B., Tims, S. R., 1966. Mathematical Analysis: An Introduction. Cambridge University Press, Cambridge.

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., 2011. Cross-situational learning: An experimental study of word-learning mechanisms. Cognitive Science 35, 480–498.

Tenenbaum, J. B., 1999. Bayesian modeling of human concept learning. In: Kearns, M. S., Solla, S. A., Cohn, D. A. (Eds.), Advances in Neural information Processing Systems 11. MIT Press, Cambridge, MA, pp. 59–65.

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

Trueswell, J. C., Medina, T. M., Hafri, A., Gleitman, L. R., 2013. Propose but verify: Fast mapping meets cross-situational word learning. Cognitive Psychology 66, 126–156.

Vlach, H. A., Johnson, S. P., 2013. Memory constraints on infants' cross-situational statistical learning. Cognition 127, 375–382.

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

A. Derivation of the main result

Here we show that Equations (1) and (3) and the statement that a word can be learnt in a finite time are equivalent. We recall that $\mathcal{K}(t)$ is the set of confounding meanings that have appeared in every episode until time t and that $\mathcal{F}_M$ comprises the first M meanings. We introduce

$$L_M(t) = \Pr\left[\mathcal{F}_M \cap \mathcal{K}(t) = \emptyset\right] \; , \tag{6}$$

the probability that the first M meanings have all failed to appear at least once in the first t episodes. From the definition of conditional probability,

$$
\begin{aligned}
L_M(t) &= \Pr\left[M \notin \mathcal{K}(t)|\mathcal{F}_{M-1} \cap \mathcal{K}(t) = \emptyset\right] \Pr\left[\mathcal{F}_{M-1} \cap \mathcal{K}(t) = \emptyset\right] \\
&= [1 - p_M(t)]L_{M-1}(t) \tag{7}
\end{aligned}
$$

where $p_M(t)$ is given by Eq. (1). This is valid for all $M > 0$ if we take $L_0(t) = 1$ for all t. Hence,

$$L_M(t) = \prod_{m=1}^{M}[1 - p_m(t)] \; . \tag{8}$$

The probability that the target meaning has been disambiguated from all other meanings is

$$L(t) = \lim_{M\to\infty} L_M(t) \equiv \prod_{m=1}^{\infty}[1 - p_m(t)] \; . \tag{9}$$

Since $p_m(t)$ is a probability, and lies between zero and one, this infinite product exists (and furthermore itself lies between zero and one).

We are interested in the case where $L^* = \lim_{t\to\infty} L(t) = 1$, since this corresponds to the word having a finite learning time. This fact follows straightforwardly from the definition of a limit, which is that for any ϵ sufficiently small, there exists a t^* such that $L(t) > 1 - \epsilon$ for all $t > t^*$. The smallest t^* for which this is true for any given ϵ corresponds to the learning time $t^*(\epsilon)$ defined in the main text.

To show that (3) is a necessary and sufficient condition for a finite learning time, our strategy is as follows. We shall first assume that (3) is true, and will then find that this implies $L^* = 1$, and the learning time is finite. We then then consider the case where (3) does not hold, and find that $L^* < 1$ as a consequence, and hence that the learning time is infinite.

27

In the first instance, where (3) is assumed, it follows that, for sufficiently large t, we have $\max_m\{p_m(t)\}$ arbitrarily small, and in particular less than unity. We can then take the logarithm of (9) and expand as a power series to obtain the convergent double series

$$\ln L(t) = -\sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{p_m(t)^k}{k} \ . \tag{10}$$

Since all terms in this double series have the same sign, the order of the summation indices can be exchanged, and the resulting double series

$$\ln J(t) = -\sum_{k=1}^{\infty}\sum_{m=1}^{\infty}\frac{p_m(t)^k}{k} \tag{11}$$

has the same limit: $L(t) = J(t)$ (Scott and Tims, 1966). (By contrast, if either of these double series diverges, then so does the other).

Using the fact that

$$\sum_{m=1}^{\infty}p_m(t)^k \leq \left(\sum_{m=1}^{\infty}p_m(t)\right)^k \tag{12}$$

we find

$$\ln L(t) = \ln J(t) \geq -\sum_{k=1}^{\infty}\frac{1}{k}\left[\sum_{m=1}^{\infty}p_m(t)\right]^k = \ln\left[1 - \sum_{m=1}^{\infty}p_m(t)\right] \tag{13}$$

where the second equality follows because it is assumed at the outset that $\sum_{m=1}^{\infty}p_m(t)$ converges to a value smaller than unity. Exponentiating both sides, we finally obtain the inequality

$$L(t) \geq 1 - \sum_{m=1}^{\infty}p_m(t) \ . \tag{14}$$

Since the condition (3) states that the series on the right-hand side of the previous expression vanishes as $t \to \infty$, we have that $L^* \geq 1$. However, L^* cannot exceed unity: hence $L^* = 1$ and the condition (3) is sufficient for a finite learning time.

We now show that (3) is also a necessary condition, i.e., that if (3) does not hold, the learning time is infinite. We consider again (11), and find that

$$\ln J(t) = -\sum_{k=1}^{\infty} \frac{1}{k} \sum_{m=1}^{\infty} p_m(t)^k \leq -\sum_{m=1}^{\infty} p_m(t) . \tag{15}$$

Now, if (3) does not hold, we must have the strict inequality

$$\lim_{t \to \infty} \ln J(t) < 0 . \tag{16}$$

If the sum for $\ln J(t)$ converges for sufficiently large t, we have that $L(t) = J(t)$, and hence $L^* < 1$ when (3) does not hold. Hence, the learning time becomes infinite at some nonzero ϵ in this case. On the other hand, if the sum for $\ln J(t)$ diverges at all finite times, so does the sum for $\ln L(t)$, which implies that $L^* = 0$, which again indicates an infinite learning time.

B. The case of logarithmic decay

In the text we discuss the situation where inference frequencies decay logarithmically, e.g., $a_m = c \ln 2 / \ln(m + 1)$, and mention that under this scenario the word is not necessarily learnt in finite time. By applying the integral test, one finds that the sum in (3) diverges at any time $t < \infty$, which in turn implies that the learning time diverges for any $\epsilon < 1$.

The way in which the learning time diverges in this example turns out to be quite subtle and somewhat revealing. Suppose we keep only the first M confounding meanings. Then, from (2) we have

$$L(t) \leq \exp\left(-[c \ln 2]^t \sum_{m=1}^{M} \frac{1}{[\ln(m + 1)]^t}\right) \tag{17}$$

since $\ln(1 - x) < -x$. Replacing the sum with an integral, and making the change of variable $u = \ln m$ we find

$$L(t) \leq \exp\left(-[c \ln 2]^t \int_{\ln(2)}^{\ln(M+2)} \frac{e^u}{u^t} du\right) . \tag{18}$$

When u is small, the exponential contribution to the integrand changes slowly, while the power-law part changes rapidly. This means that the integral is weakly dependent on M while it remains below some characteristic
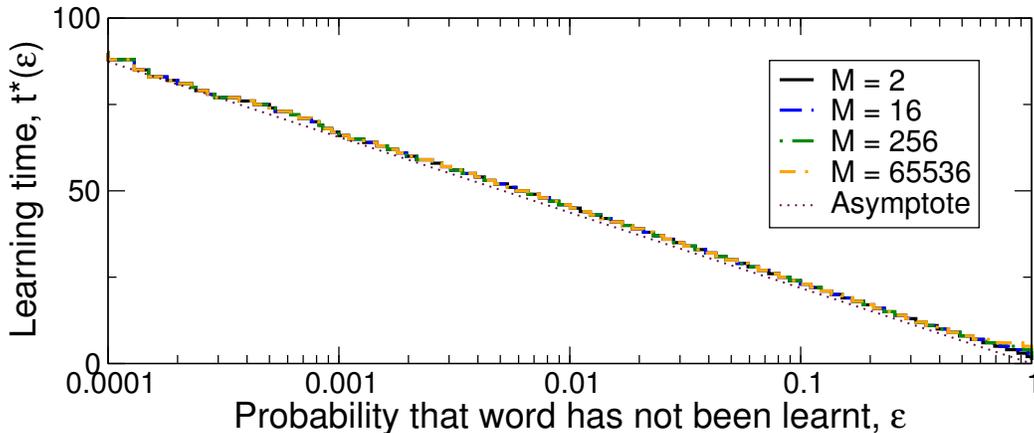
Figure 5: Time to learn a word with probability $1-\epsilon$ with confounder frequencies governed by a logarithm $a_n = c\frac{\ln(2)}{\ln(n+1)}$, with $c = 0.9$. Here the data for different sizes of the space of confounders, M, are indistinguishable except very close to $\epsilon = 1$. The learning times are close to the asymptotic result $t^*(\epsilon) \sim \ln(\epsilon)/\ln(c)$ even though formally in the limit $M \to \infty$ the learning time diverges for all ϵ.

value. In this regime, one can set $e^u = 2$ (its value at the lower end of the integral), and finds that the integral *apparently* converges to

$$\int_{\ln(2)}^{\ln(M+2)} \frac{2}{u^t} \mathrm{d}u = \frac{2}{t-1} \frac{1}{(\ln 2)^{t-1}} \left(1 - \left[\frac{\ln 2}{\ln(M+2)}\right]^{t-1}\right) \qquad (19)$$

which remains finite in the limit $M \to \infty$. From this, one can estimate the behaviour of the learning time as $\epsilon \to 0$ as $t^*(\epsilon) = \ln(\epsilon)/\ln(c)$, which as previously is governed purely by the frequency of the most common confounder. Numerical data (not shown) shown in Figure 5 correspond well with this estimate.

The integrand e^u/u^t has a minimum at $u = t$. This provides an estimate of the value of u at which we can no longer regard e^u as constant, and at which the integral begins to diverge. The estimate of t^* obtained above is therefore only valid if $u < t^*$ across the whole range of the integral, and in particular at the top end $u \approx \ln(M)$. For this model we therefore find that the learning time will be apparently finite if $M < (\epsilon)^{1/\ln(c)}$. For the case $\epsilon = 0.01$ and $c = 0.9$, we would need to consider *at least* 10^{18} confounding meanings before probing the region where the integral diverges.