Modeling the acquisition of quantifier semantics: a case study in function word learnability

Steven T. Piantadosi
Department of Brain and Cognitive Sciences, University of Rochester

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences, MIT

Noah D. Goodman
Department of Psychology, Stanford University

Abstract

This paper studies the acquisition of quantifier meanings as a case study of function word learnability. We suggest that learners construct semantic representations of quantifiers and other function words using a compositional statistical learning mechanism that operates over a small set of domain-general cognitive primitives. We present a simple cross-situational learning model that provably solves key learning problems in this domain, using a developmentally-plausible amount of data. We additionally evaluate the utility of proposed constraints on quantifier meaning, and show that learning in an unrestricted space of meanings is not substantially more difficult than learning in highly-constrained frameworks.

Introduction

In acquiring language, children discover remarkable representations that enable complex communication. They do this from seemingly impoverished evidence, to arrive at a linguistic system that goes far beyond what is directly observable in their input. This capacity is especially striking in children's acquisition of *function words* like "the," "both," and "and." Content words—which map onto objects, actions, and properties—are plausibly learned through tracking co-occurrences between words and features of the world as in cross-situational word learning models (Siskind, 1996; Vogt & Smith, 2005; Smith, Smith, Blythe, & Vogt, 2006; Yu & Ballard, 2007; Yu & Smith, 2007; Frank, Goodman, & Tenenbaum, 2007). But function words do not correspond to any plainly observable perceptual phenomena and so must be learned by other means. Function words express their meaning through semantic composition and thus embody two of the most interesting challenges of language acquisition: abstractness and compositionality.

In general, function words are a striking gap in most theories of language learning, with little to no attention from statistical approaches, and relatively incomplete and non-

computational (non-implemented) nativist theories. Even for a maximally nativist theory under which all function word meanings are innately specified, children still face a problem of mapping them to their corresponding phonetic forms. As we show, this is a substantial challenge.

Here, we study quantifiers, a subset of function words, with the aim of showing how techniques for structural statistical learning can explain learners' capacity to arrive at these representations. Informally, quantifiers specify "how much" or "how many" in utterances; formally, they are typically taken as denoting relations between sets (see also Montague, 1973; Barwise & Cooper, 1981; Keenan & Stavi, 1986; Keenan & Westerståhl, 1997). A sentence such as "most accordionists are sailors" is true if the set of accordionists who are sailors has more elements than the set of accordionists who are not. Semantically, "most" then would denote a *function* on these two sets: "most A are B" iff $|A \cap B| > |A \setminus B|$, where $\cap$ is set-intersection, $\setminus$ is set-difference and $|\cdot|$ is set-cardinality. Quantifiers may also bring presuppositional assumptions to utterances. In a sentence like "Both accordionists are sailors," the word "both" asserts that the accordionists who are sailors has cardinality two and assumes that there are exactly two relevant accordionists.

These multiple aspects of abstract (non-referential) meaning present a substantial learning challenge. However, we show that these problems can be solved by the right kind of statistical learner—one that uses Bayesian statistical inference to determine the likely representations in adults' heads that generate the observed data. The learning model we present takes naturalistic positive evidence in the form of uttered words in world contexts and infers the likely representations for each word's meaning. Our approach builds on a considerable amount of work developing rational probabilistic modeling in a *language of thought* (LOT) (Katz, Goodman, Kersting, Kemp, & Tenenbaum, 2008; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Kemp, Goodman, & Tenenbaum, 2008; Goodman, Ullman, & Tenenbaum, 2009; Ullman, Goodman, & Tenenbaum, 2010; Piantadosi, 2011; Piantadosi, Tenenbaum, & Goodman, 2012), in which learners create new conceptual systems by composing elements from a smaller set of *primitive functions*. Here, we posit that learners have access to primitive set operations like set-intersection ($\cap$) and set-difference ($\setminus$) and, when given data, learn how to compose these kinds of operations to express quantifier meanings. Learners thus might discover that a good representation for "most" is $|A \cap B| > |A \setminus B|$, instead of any other composition of set operations.

There are several specific motivations for applying a compositional LOT model here. First, semanticists often express word meanings using a structured, compositional representation system (e.g., Montague, 1973; Heim & Kratzer, 1998; Steedman, 2000), because doing so allows complex word meanings to be formalized precisely in terms of simple and well-defined logical operations. Second, as we show, a compositional representation system provides a compelling account of learning: learning consists of appropriately combining (composing) simpler logical capacities. This type of approach has been applied to several areas of language research, including lexical semantics (Siskind, 1996), number-word acquisition (Piantadosi, Tenenbaum, & Goodman, 2011), and compositional semantics (Zettlemoyer & Collins, 2005; Kwiatkowski, Goldwater, & Steedman, 2009; Piantadosi, Goodman, Ellis, & Tenenbaum, 2008). Such compositional learning is not necessarily language specific—language of thought models have also been applied to explain learning and development in other domains, including kinship relations (Katz et al., 2008), abstract relational concepts

(Kemp et al., 2008), boolean rule-based concepts (Goodman et al., 2008), intuitive notions of causality (Goodman et al., 2009), and magnetism (Ullman et al., 2010). Thus, the LOT approach is able to explain quantifier learning using only inferential and representational tools that have been independently motivated in other domains.

In all cases statistical learning over compositional hypothesis spaces provides a powerful framework for learners. We show that this kind of model is capable of learning the literal aspects of meaning and presuppositions, from positive evidence: no explicit feedback is required. This contrasts strongly with Gold-style (Gold, 1967) approaches to language learning. The key difference between our approach and Gold's is that we—like others (e.g. Horning, 1969; Chater & Vitányi, 2007; Hsu, Chater, & Vitányi, 2011)—assume utterances are drawn from the true generative distribution, not from an antagonistic teacher. Importantly, we show that an implemented version of the model is an effective learner, able to acquire quantifier meanings in several hundred to a few thousand observed utterances.

The outline of this paper is as follows: in the next section, we review previous theories of quantifier learning. We then describe what we consider to be one of the key logical challenges in learning quantifiers and other function words, the *subset problem*. Our results require a fully explicit model, so we then formalize aspects of meaning which we intend to learn—including literal and presuppositional content. We then describe the probabilistic model which can take observed utterances and discover the correct meanings in the representation system. We prove that this learning model always recovers the correct meanings and explain in detail how it solves the subset problem. We conclude with simulations showing that the learning model is computationally tractable and requires only a developmentally-plausible amount of data. The simulations also allow us to compare the utility of proposed constraints on the space of quantifier meanings.

Learnability and quantification

Quantifier learnability has previously been studied by associating natural language quantifiers with abstract devices from computability theory (van Benthem, 1984, 1986; M. Mostowski, 1998; Tiede, 1999; Florêncio, 2002; Gierasimczuk, 2007). van Benthem (1984), noted that some quantifier meanings can be computed by finite-state automata (for extensions up the automata hierarchy, see M. Mostowski, 1998). For instance, the meaning of "every" can be captured by a finite-state machine like that shown in Figure 1(a). Here, a language user wishing to check if "every A is B" would start in the double-circled state *true* and proceed to look at elements of A. Each element $a \in A$ is processed and if it is an element of B, a 1 link is followed; if it is not, a 0 link is followed. Thus, as long as every element in A is in B, the learner will stay in the *true* state; otherwise they will permanently fall into the *false* state. A similar example for "some" is shown in Figure 1(b), where one positive example is enough to change the automaton permanently to an accepting state.

This type of formalization is the basis for much previous work on quantifier learnability. Clark (1996) presents a detailed account of quantifier acquisition, providing similar automata for even more complex meanings, such as "none," "at least two," and "an even number of" (see also Clark, 2010). By formalizing meanings as finite-state automata, Clark is able to apply learnability results for regular languages (Angluin, 1987) to show that first-

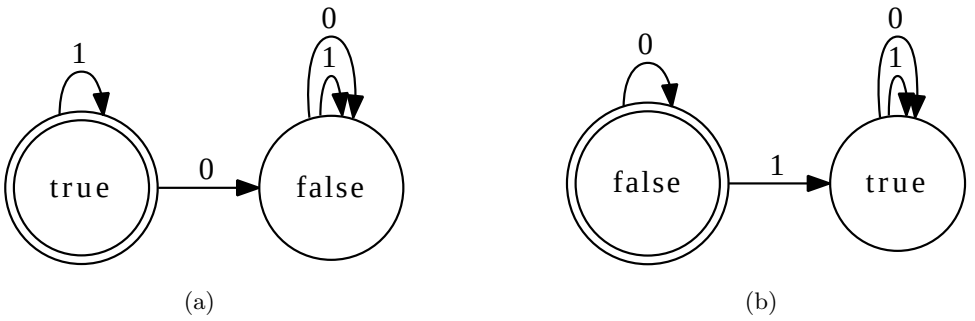

Figure 1. The representation of "every" or "all" (a) and "some" (b) in Clark (1998)'s learning model.

order quantifiers[1] can be learned jointly using positive and negative evidence. These results were extended by Tiede (1999), who showed that first-order *left increasing monotonic* quantifiers are identifiable in the limit (i.e. Gold-learnable) from *positive* evidence alone[2]. Not all first-order quantifiers are identifiable in the limit from positive evidence: Tiede shows that quantifiers that are left *decreasing* monotonic (e.g., "few"), right increasing monotonic (e.g., "several"), or right decreasing monotonic (e.g., "no") do not come with guarantees of learnability[3]. Florêncio (2002) extends these results to what he argues are psychologically plausible restrictions on learning algorithms, such as algorithms that do not care about the order of sets or only change hypotheses when they are incorrect.

While these results have mapped out the space of learnability for some quantifiers in a mathematically sophisticated way, the approach is incomplete. First, these learning theories only apply to subsets of natural language quantifiers (for instance, the left-upward-monotonic ones). A complete learning theory should handle at least everything observed in natural language. Notably lacking are theories of quantifiers which cannot be expressed in first order logic, such as "most"—what is the hypothesis space for these kinds of meanings and how does the learner navigate it? In addition, it is not clear how these learning frameworks might be extended to handle noisy evidence. In the case of quantifiers, this means perhaps incorrectly identifying the relevant sets, or occasionally hearing quantified expressions which are false. These learning theories are not implemented, meaning that it is unclear if the amount of data required is at all plausible. These theories only capture literal meanings and to our knowledge have not been extended to other aspects of meaning, such as presupposition. This is likely because they use representations which are wholly unlike anything else used in semantics and it is not exactly clear how to relate finite state ma-

[1]Those which can be expressed in first-order logic; not, e.g., "most."

[2]Left increasing monotonic quantifiers are those quantifiers Q such that $(Q\,A\,B) \rightarrow (Q\,A'\,B)$ where $A \subseteq A'$. In other words, quantifiers that, if true, can generalize to any more inclusive first set. For instance, "several" is left-upward-monotonic since if "several angry lawyers are fools" is true, then "several lawyers are fools": by increasing the size of the first set from "angry lawyers" to "lawyers," we do not make the sentence false. Note that this is not true for "few": "few angry lawyers are fools" does *not* imply "few lawyers are fools."

[3]Tiede also shows how all quantifiers with a certain form in *Presburger arithmetic* are learnable in the limit.

chines to more standard machinery in compositional semantics. Indeed, the choice of these representations seems to be primarily for mathematical convenience, rather than linguistic plausibility. Perhaps most importantly, these approaches typically do not explicitly address or solve what we see as one of the most interesting and challenging aspects of learning from positive evidence: *the subset problem.*

The subset problem in semantics

The *subset problem* is that learners may mistakenly infer (hypothesize) an under-restrictive semantic representation for a word. If this happens, it is not obvious how positive evidence could convince them they have made a mistake. As an example, SOME is logically weaker than EVERY: whenever "Every accordion is heavy," it is also true that "Some accordion is heavy," assuming there is one accordion present. If a child incorrectly guesses that "every" has the denotation of SOME, they would never receive an error signal of observing "every" used in an unexpected situation (such as where it is false). In other words, under a simple picture of learning where children only change representations when their hypothesized meaning is observed to be false, a child who thinks that "every" meant SOME will never be wrong and so may in such a framework never change their mind.

The subset problem appears in many areas of language acquisition including syntax (Wexler & Manzini, 1987; Berwick, 1985), phonology (Smolensky, 1996; Hale & Reiss, 2003), and learning compositional sentences structures (Crain, Ni, & Conway, 1994). One proposed solution is the *subset principle*, which proposes that learners have a strong innate bias for logically stronger hypotheses (Wexler & Manzini, 1987; Berwick, 1985; Smolensky, 1996; Crain, 1992, 1993; Crain & Philip, 1993; Gualmini & Schwarz, 2009; Crain & Thornton, 2000; Crain et al., 1994; Musolino, 2006). Positive evidence then compels learners to move to logically weaker hypotheses. In the case of "every" and "some," learners' initial state would be to innately prefer the meaning EVERY for both words, and then uses in other context would eventually show them that "some" has a logically weaker meaning.

The subset principle requires learners to have a very specific initial state with meanings or hypotheses ordered by logical strength. Unfortunately, there are several problems with accounts based on the subset principle. Even within quantifiers, it is not always possible to order hypotheses by logical strength, as with "most" and "many." Additionally, an innate subset principle seems much less plausible when one considers that the subset problem is faced more generally in lexical acquisition. It is hard to imagine that learners have an innate specification that PUSHING is more specific than TOUCHING, GREYHOUND is more specific than DOG, etc. Even worse, the subset principle appears unable to handle noisy input because it implicitly formalizes an irreversible process. Once learners come to think that a word means SOME, no (positive) evidence can ever convince them that it really meant EVERY. A related problem, the "triggering problem" (Borer & Wexler, 1987), is that the subset-principle accounts fail to explain the fact that learners do not *immediately* change hypothesized meanings when given contrary evidence, as many subset-principle accounts seem to assume. Though one can imagine versions of the subset principle to solve both of these problems—perhaps learners have some threshold for the amount of evidence required to change their meaning—to our knowledge such a system has never been formalized or shown to learn correctly.

The subset principle also appears to make incorrect predictions about learning tra-

jectories. Musolino (2006) reviews predictions of the semantic subset principle that fail in behavioral tests. He presents examples involving compositional sentences structures, in which an ambiguous sentence has multiple interpretations that stand in subset relationships, but children do not make the most specific generalization possible early in learning[4]. Similarly, in word learning, Xu and Tenenbaum (2007) showed that children's generalization of word labels is not maximally specific, but instead follows the predictions of a Bayesian statistical learner.

In sum, there are compelling problems with the subset principle in language acquisition. The subset principle requires a complex set of language-specific hypotheses and computations, is not clearly able to handle noisy data or all quantifier meanings and appears to make already falsified predictions. Moreover no implemented models exist demonstrating the computational tractability and theoretical soundness of these proposals. Despite these problems, the subset principle does capture the appealing intuition that more specific hypotheses should be preferred when they are right. A learner should disprefer for "every" to mean SOME because it makes the incorrect, broad prediction that "every" should be a possible option in all situations where SOME is.

We formalize this intuition in a probabilistic model, drawing on domain general mechanisms of correct statistical inference to solve the subset problem. We expect that the approach to the subset problem taken here is considerably more general and applicable to subset problems in areas like syntax and phonology as well. In the next section, we first describe the space of quantifier meanings we consider and then describe the statistical learning model.

Aspects of quantifier meaning

We begin our approach to quantifier learning by describing the types of representations that learners must eventually acquire. We are motivated by contemporary semantic theories since these constitute our "best guess" for adults' knowledge of quantifier meanings. However, our implementation is a simplification that is intended primarily to address the interesting challenge of learning both *literal meaning* and *presupposition* for a realistic set of quantifiers. After discussing these aspects of meaning, we present the target meanings and hypothesis space for learners.

Literal meaning

We follow Heim and Kratzer (1998) in supposing that to a first approximation, the literal meaning of quantifiers can be captured with *generalized quantifiers*, logical operations that denote relations between sets (see also Montague, 1973; Barwise & Cooper, 1981; Keenan & Stavi, 1986; Keenan & Westerståhl, 1997). For instance, the sentence "Some reporter is a liar" might be mapped to a logical representation,

$$(\textit{nonempty?}\ (\textit{intersection reporters liars})). \tag{1}$$

[4]For instance, sentences such as "Every student can't afford a new car." could mean either (i) for each student s, s cannot afford a new card, or (ii) it is not the case that every study can afford a new car. Since (i) implies (ii), the subset principle implies early learners should interpret the sentences as (i), not (ii); but, the opposite is true (Musolino, Crain, & Thornton, 2000).

Throughout this paper we use *prefix notation*, meaning that a function f applied to an argument x is written $(f\ x)$. Expression (1) is an expression that says that the intersection of the set of reporters and liars is not empty. It is built using two logical operations: *intersection* computes the set-intersection of its arguments, and *nonempty?* checks if a set is not empty[5]. This expression refers to two sets, the set of *reporters* and the set of *liars* (each in the relevant context). To arrive at this representation, comprehenders would use their compositional semantics, composing the individual meanings of words in the sentence. Simple formalized systems for this type of language processing can be found in Blackburn and Bos (2005); detailed linguistic accounts can be found in Heim and Kratzer (1998) and Steedman (2000)[6]. To a first approximation, most systems would map "reporters" to the set of reporters, "liar" would map to the set of liars, and "some" would have a special denotation, a function of two sets:

$$\lambda\ A\ B\ .\ (\textit{nonempty?}\ (\textit{intersection}\ A\ B)). \tag{2}$$

This notation, lambda calculus, provides a convenient formalism for expressing functions. Here, "$\lambda\ A\ B\ .$" denotes that the expression (**??**) is a function of the variables A and B which gives a return value consisting of everything after the ".". The compositional semantics of English would have to pass *reporter* for "reporter" as the argument A and *lied* as the argument B in order to arrive at (1). Many quantifier meanings can be written down as lambda expressions like this that take two sets and return a truth value. For instance, "every" might be denoted

$$\lambda\ A\ B\ .\ (\textit{subset?}\ A\ B) \tag{3}$$

where *subset* is a function which is true if the first set is a subset of the second. "No" (or "none of the") might be written as

$$\lambda\ A\ B\ .\ (\textit{empty?}\ (\textit{intersection}\ A\ B)). \tag{4}$$

We note that we could have written down each of the above quantifiers in first-order logic, using $\forall$ and $\exists$. The use of set-theoretic operations is motivated by other quantifiers meanings which provably cannot be expressed in first-order logic (A. Mostowski, 1957; Barwise & Cooper, 1981). For instance "most" cannot be written down using $\exists$ and $\forall$, intuitively because "most" requires comparing potentially arbitrarily large cardinalities, which is impossible in first order logic. However, "most" can be expressed in our notation by assuming operations for cardinality comparison:

$$\lambda\ A\ B\ .\ (\textit{card}{>}\ (\textit{intersection}\ A\ B)\ (\textit{set-difference}\ A\ B)). \tag{5}$$

Here, "*card*>" is a function that compares the cardinality of its first argument to the cardinality of its second. Note that for "most" and all the other quantifiers studied here, there are many equivalent ways of writing their meanings. Alternative formalizations, when

[5] In standard set-theoretic notation this would be $\textit{reporters} \cap \textit{liars} \neq \emptyset$; in first-order logical notation it might be written $\exists x.\textit{reporter}(x) \wedge \textit{lied}(x)$. We use prefix notation keeping in line with previous work (e.g Piantadosi et al., 2011; Piantadosi, Tenenbaum, & Goodman, 2009).

[6] Here, we will focus only on the meaning of the quantifier and not how it compositionally combines with other words, though see Piantadosi et al. (2008) and Zettlemoyer and Collins (2005) for theories of learning compositional structures.

treated as explicit theories of the computational processes underlying these word meanings have been argued to give rise to different behavioral hallmarks (Hackl, 2009; Pietroski, Lidz, Hunter, & Halberda, 2009), but these distinctions will not be addressed in this work.

Presupposition

The second aspect of quantifier meaning is *presupposition*, which captures the assumptions that are required for a statement to receive a truth value (see Heim & Kratzer, 1998, section 6.7, for an overview). Our primary representational choices build off proposals in semantics and philosophy of language dating back to Russell (1905, 1957) and Strawson (1950)[7], who argued about the correct way to handle presuppositions in the definite determiner, "the." Russell argued that the meaning of sentences like "The A is B" asserts that A is true of exactly one element and that element is in B. In other words "The accordionist is cooking" is true if and only if there is exactly one accordionist and that accordionist is cooking. This proposal captures the notion that "the" can only be used in situations where there is a unique referent. However, as argued by Strawson (1950), this account is lacking in that it seems to assign truth values to sentences which intuitively may not even have truth values. Strawson's sentence, "The present king of France is bald" would be strictly false under Russell's account, since it is not true that there is exactly one present king of France. Strawson argues that our intuitions really say this sentence *does not have* a truth value (see also Russell, 1957; Von Fintel, 2004). Strawson argues that sentences like "The present king of France ... " *presuppose* the existence of a king of France, rather than assert it. In order for such sentences to be true or false, there must exist a king of France. If there is no king of France, the sentence is neither true nor false. Indeed, violations of such background assumptions appear to have different behavioral hallmarks than truth-value violations of asserting something false (Langford & Holmes, 1979).

Presuppositions are an important aspect to quantifier meanings. For instance, in a situation where there is exactly one sailor, it is bizarre to assert

"Both sailors are happy." (6)

regardless of whether the one sailor is happy. Intuitively, sentence (6) requires as part of its background assumption that there are exactly two sailors and it is difficult to say whether it is strictly true or false if this assumption is not satisfied.

We capture presuppositional aspects of meaning by assuming that semantic representations have two parts: the presupposed content and the asserted content (see Karttunen & Peters, 1979; Heim, 1991). For instance, "both" would presuppose exactly two elements in A and assert that A is a subset of B:

Presupposed	$\lambda\ A\ B\ .\ (\mathit{doubleton?}\ A)$
Asserted	$\lambda\ A\ B\ .\ (\mathit{subset}\ A\ B)$

[7]We do not wish to get bogged down in the details of the semantic analyses of these words, or the large philosophical and linguistic literature devoted to more thoroughly developing theories of semantics, reference, and presupposition; for a detailed description, see (see Ludlow & Neale, 2008).

Word	Presupposition	Literal meaning
the	$\lambda\ A\ B\ .\ (singleton?\ A)$	$\lambda\ A\ B\ .\ (nonempty?\ (intersection\ A\ B))$
a/some	$\lambda\ A\ B\ .\ TRUE$	$\lambda\ A\ B\ .\ (nonempty?\ (intersection\ A\ B))$
one	$\lambda\ A\ B\ .\ (nonempty?\ A)$	$\lambda\ A\ B\ .\ (singleton?\ (intersection\ A\ B))$
two	$\lambda\ A\ B\ .\ (nonempty?\ A)$	$\lambda\ A\ B\ .\ (doubleton?\ (intersection\ A\ B))$
three	$\lambda\ A\ B\ .\ (nonempty?\ A)$	$\lambda\ A\ B\ .\ (tripleton?\ (intersection\ A\ B))$
both	$\lambda\ A\ B\ .\ (doubleton?\ A)$	$\lambda\ A\ B\ .\ (doubleton?\ (intersection\ A\ B))$
either	$\lambda\ A\ B\ .\ (doubleton?\ A)$	$\lambda\ A\ B\ .\ (singleton?\ (intersection\ A\ B))$
neither	$\lambda\ A\ B\ .\ (doubleton?\ A)$	$\lambda\ A\ B\ .\ (empty?\ (intersection\ A\ B))$
every	$\lambda\ A\ B\ .\ (nonempty?\ A)$	$\lambda\ A\ B\ .\ (subset?\ A\ B)$
most	$\lambda\ A\ B\ .\ (nonempty?\ A)$	$\lambda\ A\ B\ .\ (card{>}\ (intersection\ A\ B)\ (set\text{-}difference\ A\ B))$
none/no	$\lambda\ A\ B\ .\ (nonempty?\ A)$	$\lambda\ A\ B\ .\ (empty?\ (intersection\ A\ B))$

Figure 2. Target quantifier meanings for the learning model.

Here, *doubleton?* is a function which is true if given a set of size two. In principle these two aspects of meaning could be combined within one single representation:

$$\lambda\ A\ B\ .\ (presup\ (doubleton?\ A)\ (subset\ A\ B), \tag{7}$$

where *presup* is a function which returns undefined if its first argument is false and it returns its second argument if the first argument is true.

The learning setup

Given the above aspects of meaning, we can define a *lexicon* to be a mapping from words to literal meanings and presuppositions. In order to provide data for the learning model, we must define the adult "target" lexicon for learning. This target lexicon is shown in Figure 2. The particular meanings in this lexicon are *not* meant to provide an exact (and uncontroversial) account of what these words mean in English—that is one as-of-yet unaccomplished goal of modern semantics. Instead, these particular meanings are meant only to capture some interesting aspects of English quantifier semantics and present an approximation to the task that learners of English face, for the purposes of implementing a learning model.

The target lexicon is used to label sets of objects with words that an adult would be likely to say. The learning situation is schematically shown in Figure 3. Here there is a collection of objects which are shown in a Venn diagram with the sets A and B. The adult speaker has the correct target lexicon (Figure 2) in their head and uses it to generate a word label in the context of an observed instance of objects, following a pragmatic model we describe later. If all the As are Bs, for instance, the adult may be likely to utter "every." From the learner's perspective, all that is available is the uttered word "every" and the sets in the context. Their job is to take a collection of these sets and noisy adult labels and infer the meaning of each word. Thus, the meanings in the target lexicon are *not* directly provided to the learner—the learner must infer them from observed usage in context.

From the learner's point of view, the space of possible representations is much larger than those in Figure 2. To formalize a plausible space of meanings, we first suppose that the learner considers all possible quantifiers that can be generated from a context-free *grammar*

of concepts (Goodman et al., 2008). The intuitive appeal of a grammar is that it requires very little specific knowledge to be "built in"—it requires learners to know only the primitive functions and how to compose them. From that, learners can build a vast set of possible representations, potentially ones of considerable computational power (see Piantadosi et al., 2012).

The specific grammar we use for learner's hypotheses is shown in Figure 4. The grammar includes a number of primitive operations that manipulate sets (*union*, *intersection*, *set-difference*), small-set cardinalities (*singleton?*, *doubleton?*, *tripleton?*), and which can relate properties of sets to truth values (*subset?*, *empty?*, *nonempty?*, *exhaustive?*). We also include the ability to form trivial expressions such as $\lambda\ A\ B\ .\ true$. We note that while this grammar includes primitives which are not strictly logically necessary for our model, we aim to describe a set of general conceptual resources that are also useful in other areas of learning, where we have used similar grammars (Piantadosi et al., 2012).

To see how the grammar could generate an expression such as,

$$\lambda\ A\ B\ .\ (singleton?\ (union\ A\ B)), \tag{8}$$

we first begin with the *START* symbol. We then recursively expand nonterminals according to the possible rules in Figure 4 until no more nonterminals remain. The only possible way to expand *START* is to $\lambda\ A\ B\ .\ BOOL$, meaning we always will generate an expression representing a function of two arguments, A and B. This function returns a boolean (*BOOL*) since *BOOL* only expands to functions which return boolean values. For instance, a *BOOL* can expand to $(singleton?\ SET)$, yielding the expression $\lambda\ A\ B\ .\ (singleton?\ SET)$. Next, we may expand *SET* to $(union\ SET\ SET)$, yielding $\lambda\ A\ B.\ (singleton?\ (union\ SET\ SET))$.

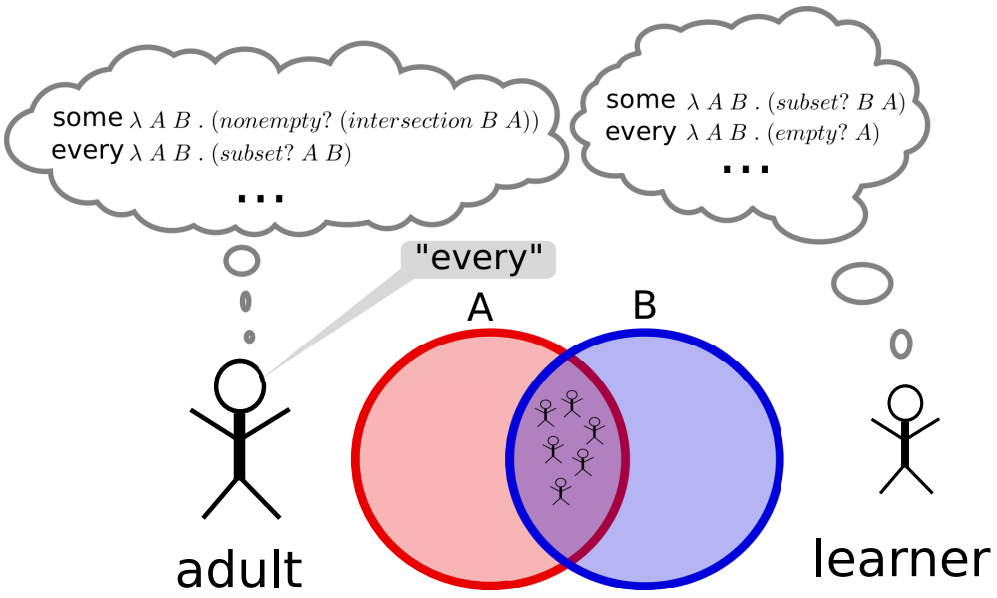

Figure 3. The learning setup for the model. The learner observes two sets A and B, and hears the parent utter a word. The word is generated according to the correct adult lexicon. The learner must use observed word usages in context to infer likely word meanings.

Nonterminal		Expansion	Gloss
START	→	$\lambda\ A\ B\ .\ BOOL$	Function of A and B
BOOL	→	$true$	Always true
	→	$false$	Always false
	→	$(card{>}\ SET\ SET)$	Compare cardinalities (>)
	→	$(card{=}\ SET\ SET)$	Check if cardinalities are equal
	→	$(subset?\ SET\ SET)$	Is a subset?
	→	$(empty?\ SET)$	Is a set empty?
	→	$(nonempty?\ SET)$	Is a set not empty?
	→	$(exhaustive?\ SET)$	Is the set the entire set in the context?
	→	$(singleton?\ SET)$	Contains 1 element?
	→	$(doubleton?\ SET)$	Contains 2 elements?
	→	$(tripleton?\ SET)$	Contains 3 elements?
SET	→	$(union\ SET\ SET)$	Union of sets
	→	$(intersection\ SET\ SET)$	Intersection of sets
	→	$(set\text{-}difference\ SET\ SET)$	Difference of sets
	→	A	Argument A
	→	B	Argument B

Figure 4. A grammar that generates quantifier meanings.

If we then expand the first SET to A and the second set to B, we will generate expression (8).

Of course, one can also generate all of the hypotheses shown in Figure 2. In general, there are a vast number of potential hypotheses that can be generated according to this grammar. Some of these are relatively complex. For instance,

$$\lambda\ A\ B\ .\ (nonempty?\ (union\ (set\text{-}difference\ B\ A)\ (set\text{-}difference\ A\ B))) \tag{9}$$

is a quantifier literal meaning or presupposition which could be expressed in this language. This is true iff there are As that are not Bs, *or* there are Bs that are not As. The challenge for the learner, then, is to take data (words in context) generated according to the target meaning in Figure 2 and find likely hypotheses generated by this grammar for *each* word meaning.

The probabilistic model

We have described a learning setup where adults have a target set of meanings and they utter words in the context of two sets. The learner has a grammar for generating hypotheses and must take the observations (sets and words) and infer the likely meaning in adults' heads that generated the observed data. The goal of the probabilistic model is to solve this problem: determine the best representation (generated from the grammar) given some observations of adult utterances in context (Figure 3). The probabilistic model we present can handle all the challenges we have described—noisy evidence, subset relationship, complex meanings with literal and presuppositional content, and a large space of logically possible hypotheses. While we present the model specifically in the context of quantifier acquisition, it is actually considerably more general to learning semantic or syntactic structures.

Let $w_1, w_2, \ldots, w_k$ be the words that the learner is trying to discover meanings for. We denote the meaning of word w_i by m_i, and the collection of all meanings as

$m = (m_1, m_2, \ldots, m_k)$. In the case of quantifiers, m_5 might be the literal meaning and presupposition for the word w_5 = "every"[8]. It turns out to be important that we simultaneously learn *sets* of words, rather than only individual ones: we will show that the solution to the subset problem requires learners to at least implicitly consider alternatives that the speaker could have uttered.

Assume that the learner hears a sequence of uttered words $u_1, u_2, \ldots, u_n$ (with $u_i \in \{w_1, w_2, \ldots, w_k\}$) each in a corresponding context $c_1, c_2, \ldots, c_n$. For the purposes of quantifiers, we will take each c_i as specifying all information about the relevant sets. So if a parent says, "most crocodiles are hungry," we can consider c_i = "_____ crocodiles are hungry", where the learner knows that in this context, the first set A is the set of crocodiles and the second set B is the set of things that are hungry. Learning this semantic compositionality of English is an interesting challenge, but it is not tackled here. Instead, we focus on how learners figure out the set relations, given knowledge of the relevant contextual and syntactic information.

We are interested in computing $P(m \mid u_1, \ldots, u_n, c_1, \ldots, c_n)$, the probability of a set of meanings m, given observed contexts and utterances. By Bayes rule,

$$P(m \mid u_1, \ldots, u_n, c_1, \ldots, c_n) \propto P(u_1, \ldots, u_n \mid m, c_1, \ldots, c_n) \cdot P(m). \tag{10}$$

The left hand side of this equation, $P(m \mid u_1, \ldots, u_n, c_1, \ldots, c_n)$ is what a learner figures out—given some utterances $u_1, \ldots, u_n$ in contexts $c_1, \ldots, c_n$, what is the probability of any *particular* hypothesized set of meanings m. By Bayes rule, this is written as the product of two terms. First, $P(m)$ is a prior probability on meanings, which we construct by converting the grammar in Figure 4 into a probabilistic context free grammar that biases learners to prefer *simple* (short) expressions in their LOT. The second term, $P(u_1, \ldots, u_n \mid m, c_i, \ldots, c_n)$, is the *likelihood*. The likelihood measures the probability that $u_1, \ldots, u_n$ would be produced in their corresponding contexts if m was the correct meaning. This is the probability that in the contexts $c_1, \ldots, c_n$, an adult would have uttered $u_1, \ldots, u_n$ *if* they had m as their meaning. This term is essentially the learner's view of language production—given a hypothetical set of meanings and observed world contexts, how likely would the adult have been to utter each meaning.

In the likelihood, it makes sense to assume that each utterance u_i depends on m and c_i, and is independent of the other utterances and contexts once m and c_i are known: what you say (u_i) depends only on your set of meanings m and the current context (c_i). This means that the likelihood can be rewritten as

$$P(u_1, \ldots, u_n \mid m, c_i, \ldots, c_n) = \prod_{i=1}^{n} P(u_i \mid m, c_i). \tag{11}$$

There is some subtlety in constructing a likelihood that is statistically valid and leads to effective learning. In particular, when there are both presuppositions and assertions, both of these parts of the utterance must affect production. Our implementation is *Gricean*: it assumes primarily that speakers tend to say things which are true and relevant to the current context (Grice, 1975).

[8]The meanings need not necessarily be semantic—they could also include pieces of syntactic structure as in, for instance, Combinatory Categorial Grammar (Steedman, 2000).

Formally, we will assume that each utterance (word) u has a "weight" (unnormalized production probability) $w(u)$ which depends on the utterance's informativeness. In particular, we assume that in choosing among relevant and true words, speakers prefer to utter ones which are true less often since they provide more specific information. This is why, for instance, it is infelicitous to say "a" in a context where "the" is also true[9]. Recent computational pragmatics modeling along these lines can be found in (Frank & Goodman, 2012; Bergen, Goodman, & Levy, 2012; Goodman & Stuhlmüller, 2012). Formally, we assume that in a context c_i, an utterance u_i has a "weight" $w(u_i)$ given by,

$$w(u_i) = \frac{1}{\nu + p_t(u_i)}, \tag{12}$$

where $p_t(u_i)$ is the probability that u_i is true in an a typical (average) context and $\nu = 0.1$ is a constant "smoothing" term that prevents the weights from becoming too large. The form of this equation is only for convenience and is meant only to capture the fact utterances which are more rarely true (high $p_t(u_i)$) should be more likely to be uttered when they are true.

The full likelihood model works as follows: we suppose that speakers first choose with some probability α_p whether or not to say a presuppositionally-valid utterance. Assuming they do, they choose from the literally true (and presuppositionally-valid) utterances with some probability α_t and some random presuppositionally-valid utterance otherwise. In each case, once a speaker has decided the truth- and presuppositional-value of the utterance they say, they sample from the appropriate words with a probability given by (12) above. Thus, if α_p and α_t are high, this model essentially assumes that speakers will tend to say true, presuppositionally-valid utterances and sample them according to a measure of informativeness. The amount of noise in the data is characterized by α_p and α_t, with the former quantifying how often speakers make pragmatic violations and the latter specifying how often they make truth-value violations[10].

Note that for each utterance, learners are *not* told whether the utterances was true and/or presuppositionally-valid. Instead, they only know that each utterance was generated by a speaker who cared about both literal truth and presuppposition. Because of this, the learner's estimate of the probability of an utterance, $P(u_i, \mid m, c_i)$, must take into account all of the ways that it could have been generated (i.e. either as a true and pragmatically-felicitous utterance, a false and pragmatically-felicitous utterance, or as a pragmatically-infelicitous utterance). If u_i is true and presuppositionally valid in context c_i, then

$$P(u_i \mid m, c_i) = \frac{\alpha_p \cdot \alpha_t \cdot w(u_i)}{W_{p\wedge t}(c_i)} + \frac{\alpha_p \cdot (1-\alpha_t) \cdot w(u_i)}{W_p(c_i)} + \frac{(1-\alpha_p) \cdot w(u_i)}{W(c_i)}. \tag{13}$$

Here, $w(u_i)$ is the weight of the quantifier u_i, given by (12). $W_{p\wedge t}(c_i)$ is the sum of the weights of all true and presuppositionally valid words for c_i, $W_p(c_i)$ is the sum of all weights

[9]Closely related linguistic accounts can be found in Heim (1991), who proposed explaining these intuitions—and others that are unrelated to quantification—with a pragmatic principle, *maximize presupposition*: all else being equal, speakers should prefer utterances with the strongest presuppositions (see also Sauerland, 2003; Schlenker, 2006; Singh, 2009).

[10]A key here is that presuppositional violations and truth-value violations lead to different utterance probabilities. Without this condition, learners would not be able to separate the presuppositional and literal aspects of meaning.

of presuppositionally valid words for c_i, and $W(c_i)$ is the sum of weights of all words. So, for instance, the second term is included because a word could have been generated by choosing a word at random from the presuppositionally valid words, ignoring truth values. This happens with probability $\alpha_p \cdot (1-\alpha_t)$ and generates the word u_i with probability $w(u_i)/W_p(c_i)$. To score the probability of words which are either false or presuppositionally invalid, the corresponding terms from (13) are dropped. For instance, if u_i is not presuppositionally valid, only the last term in (13) is included since the word could not have been generated by choosing from presuppositionally valid words, or true words. This likelihood function is meant only to capture the tendency to utter true and presuppositionally valid words; our learnability results do not depend on its specific form, only on the fact that it specifies a valid generative model.

Equation (13) implements the size principle (Tenenbaum, 1999), an important feature of Bayesian statistical learning models that has support form theory and experiments (e.g. Xu & Tenenbaum, 2007; Piantadosi et al., 2008; Frank et al., 2007; Piantadosi et al., 2011). For us, the size principle holds that the probability that any particular word u_i is used in c_i depends on the weight of the words which alternatively could have been uttered[11]. This consideration of alternatives allows learners to solve the subset problem in quantifier learning and language learning more generally.

How the size principle solves the subset problem

To illustrate how the size principle solves the subset problem, it is useful to consider the example of "every" and "some," and suppose that they are the only words in the lexicon. For simplicity, in this section, we also assume that $\alpha_p = \alpha_t = 1$, so that the only utterances under consideration are true and presuppositionally valid. In this case, the likelihood $P(u_i \mid m, c_i)$ is only the first term of (13), which reduces to

$$P(u_i \mid m, c_i) = \frac{w(u_i)}{W_{p\wedge t}(c_i)}. \tag{14}$$

As above, the subset problem for these words is that learners might incorrectly believe that "every" meant SOME and would never receive evidence contradicting this, since whenever "every" is true, SOME is also true. In the size principle formulation, the key is to look at the likelihood of observed instances of "some" when "every" means SOME, compared to when "every" means EVERY. If "every" meant EVERY, then most of the time it would not be true in a context where "some" was uttered, since "every" is logically stronger. This means that it will typically be the case that

$$P(\text{“some”} \mid m, c_i) = \frac{w(\text{“some”})}{w(\text{“some”})} = 1. \tag{15}$$

since if "every" is not true, $W_{p\wedge t}(c_i)$ is only the weight of "some."

In contrast, if "every" meant SOME both would be true in the same situations and the denominator $W_{p\wedge t}(c_i)$ in (14) would increase, making the observed instances of "some" *less likely*:

$$P(\text{“some”} \mid m, c_i) = \frac{w(\text{“some”})}{w(\text{“some”}) + w(\text{“every”})} < 1. \tag{16}$$

[11]This is why in (11), the probability of u_i depends on m and not just m_i.

Letting "every" mean SOME decreases the likelihood of the observed instances of "some." The reason for this is intuitive: if "every" meant SOME, each instance of "some" would have to have been sampled from two possible true utterances rather than just one. Analogously, in a sequence of coin flips, ten heads in a row are more likely under a hypothesis of a coin with heads on both sides, than a coin with heads on one side and tails on another. Intuitively if both heads and tails were possible, the sequence of heads is less likely to occur; if "every" and "some" could both be used in many contexts, the observed instances of "some" would have to be less likely. Probability mass should not be held out for events that don't occur. This is an application of the size principle (Tenenbaum, 1999): hypotheses can assign the observed utterances higher likelihood if they predict that fewer words are true in each context.

The size principle is similar to the subset principle proposed previously in that it prefers meanings which are logically strong, or true less often. However, it differs from the subset principle in the root cause of this preference. The size principle prefers meanings which are true less often because they can assign the observed utterances a higher *likelihood*, all else being equal. In contrast, the subset principle puts the bias in the *prior*, assuming that learner's innate expectations lead them to prefer stronger logical meanings. The advantage of putting the preference in the likelihood is that it falls out very naturally by positing that learners think about how language is generated. Once learners realize is that utterances are generated using a set of meanings, and that the total probability of all possible utterances must sum to 1, they can gain in the likelihood by positing that fewer words are true in each context[12].

In the next section we show that this Bayesian framework is considerably more powerful than only solving the subset problem: it can always learn the correct set of meanings.

The Bayesian model is provably learnable

This section is meant to introduce a simple proof of the learnability of meanings in a Bayesian framework. The proof is not novel—it is well-known that in the limit, the data will support the correct model if the correct model is in the hypothesis space (see Li and Vitányi (2008) for a more detailed theory and Hsu et al. (2011) for related proofs). We present it here because we hope to refine the debate on learnability, moving away from questions of what is in principle learnable, to questions of what can be learned by plausible computational models on realistic data. To show learnability with this setup, we will consider the *Bayes factor*, a measurement which quantifies the strength of belief an ideal learner should have for one model over another (Jeffreys, 1961). The Bayes factor is defined to be the log ratio of the posterior probabilities of two statistical models. In this case, one statistical model will be data generated with the correct set of meanings, $\hat{m}$. The alternative model will be any other set of meanings, m. The Bayes factor in favor of $\hat{m}$ is then given by

$$BF = \log \frac{P(\hat{m} \mid u_1, \ldots, u_n, c_1, \ldots, c_n)}{P(m \mid u_1, \ldots, u_n, c_1, \ldots, c_n)}. \tag{17}$$

[12]In this sense, the size principle is a simple consequence of formalizing a fully generative statistical model. One could imagine alternative models that, for instance, set $P(u_i \mid m, c_i) = \alpha$ if u_i is true, and $1 - \alpha$ if u_i is false. Such a model is intuitive in penalizing incorrect meanings, but doesn't specify a valid probability distribution and would fail to solve the subset problem.

The Bayes factor ranges from negative infinity (definitive support of m) to positive infinity (definitive support of $\hat{m}$) and equals zero when $\hat{m}$ and m have the same posterior probability (the data favors neither). We will show that as the amount of data gets large, the Bayes factor in support of the correct model over any alternative goes to infinity with probability 1. Thus, with enough positive examples, learners will accumulate an arbitrarily large amount of evidence supporting the correct set of meanings.

Using Bayes rule (10), we can rewrite the Bayes factor as

$$\log \frac{P(\hat{m})P(u_1, \ldots, u_n \mid \hat{m}, c_1, \ldots, c_n)}{P(m)P(u_1, \ldots, u_n \mid m, c_1, \ldots, c_n)}. \tag{18}$$

As above, we assume that each u_i depends only on c_i and is conditionally independent of all other u_j and c_j ($j \neq i$). In other words, each utterance depends only on the context it occurs in and not any other utterances or contexts. This means that we can factor (18), as

$$\log \left[\frac{P(\hat{m})}{P(m)} \prod_{i=1}^{n} \frac{P(u_i \mid \hat{m}, c_i)}{P(u_i \mid m, c_i)} \right], \tag{19}$$

which can be re-written to

$$\log \frac{P(\hat{m})}{P(m)} + \sum_{i=1}^{n} \log \frac{P(u_i \mid \hat{m}, c_i)}{P(u_i \mid m, c_i)}. \tag{20}$$

This says that the Bayes factor can be re-written as the sum of the log ratio between the prior on m and $\hat{m}$, a constant, plus the sum of the ratio between the likelihoods on each data point. We are concerned with what happens for typical learners who get increasing amounts of potentially noisy data generated from the correct target grammar (i.e. adult speech). For this, we can compute the expected Bayes factor (20) after observing n data points:

$$\mathop{\mathbb{E}}_{\substack{u_1, u_2, \ldots \\ c_1, c_2, \ldots}} \left[\log \frac{P(\hat{m})}{P(m)} + \sum_{i=1}^{n} \log \frac{P(u_i \mid \hat{m}, c_i)}{P(u_i \mid m, c_i)} \right] \tag{21}$$

where the expectation is taken over all sequences of utterances u_i and contexts c_i. Equation (21) can be simplified:

$$\begin{aligned} &\log \frac{P(\hat{m})}{P(m)} + \mathop{\mathbb{E}}_{\substack{u_1, u_2, \ldots \\ c_1, c_2, \ldots}} \left[\sum_{i=1}^{n} \log \frac{P(u_i \mid \hat{m}, c_i)}{P(u_i \mid m, c_i)} \right] \\ &= \log \frac{P(\hat{m})}{P(m)} + \sum_{i=1}^{n} \mathop{\mathbb{E}}_{u_i, c_i} \left[\log \frac{P(u_i \mid \hat{m}, c_i)}{P(u_i \mid m, c_i)} \right] \\ &= \log \frac{P(\hat{m})}{P(m)} + n \cdot \mathop{\mathbb{E}}_{u_i, c_i} \left[\log \frac{P(u_i \mid \hat{m}, c_i)}{P(u_i \mid m, c_i)} \right]. \end{aligned} \tag{22}$$

This implies that if

$$\mathop{\mathbb{E}}_{u_i, c_i} \left[\log \frac{P(u_i \mid \hat{m}, c_i)}{P(u_i \mid m, c_i)} \right] > 0, \tag{23}$$

the expected Bayes factor will increase without bound as n gets large—as more and more data points are observed, the correct system $\hat{m}$ will come to be favored over any alternative. This will eventually overwhelm any effect of the prior log ratio $\log \frac{P(\hat{m})}{P(m)}$, meaning that learners will eventually assign $\hat{m}$ the highest posterior probability.

To show that (23) holds, note that because we assume utterances are generated from the correct adult grammar via $P(u_i \mid \hat{m}, c_i)$,

$$\mathop{\mathbb{E}}_{u_i,c_i}\left[\log \frac{P(u_i \mid \hat{m}, c_i)}{P(u_i \mid m, c_i)}\right] = \mathop{\mathbb{E}}_{c_i}\left[\sum_{u_i} P(u_i \mid \hat{m}, c_i) \log \frac{P(u_i \mid \hat{m}, c_i)}{P(u_i \mid m, c_i)}\right]. \tag{24}$$

A standard theorem in information theory and probability, known as the *Gibbs inequality*, holds that

$$\sum_x A(x) \log \frac{A(x)}{B(x)} > 0 \tag{25}$$

if A and B are different distributions on elements x. A proof of this is provided in Cover and Thomas (2006, Theorem 2.6.3). This applies to (24), by letting $A(u_i) = P(u_i \mid \hat{m}, c_i)$ and $B(u_i) = P(u_i \mid m, c_i)$. Thus, for any c_i the term $\sum_{u_i} P(u_i \mid \hat{m}, c_i) \log \frac{P(u_i|\hat{m},c_i)}{P(u_i|m,c_i)} > 0$ meaning that (24) must be greater than zero (i.e. 23 holds). Thus, (22) tends to infinity as n increases. In expectation, an ideal learner will favor $\hat{m}$ over any alternative m with enough data, since each data provides, on average, evidence in favor of the correct meanings.

Note that we have made no assumptions about the form of $P(u_i \mid m, c_i)$—that is, about *how* the set of meanings m give rise to utterances. We have not even assumed the form we use in our implemented model (e.g. (13)). Under *any* such system, corresponding to any linguistic system, the above argument will hold. Importantly, we have *not* assumed that the learner hears negative evidence, or perfect data—these are not required for learning. Rather, the Bayesian setup motivates an importantly different requirement: learners should understand how adults *would* speak for any given set of semantic representations. It is then possible for an ideal learner to figure out the word meanings using standard probabilistic inference.

The implemented learning model

More important than establishing learnability in theory is showing that the correct meanings are learnable with a developmentally-plausible amount of data. Here, we use an implemented version of the model to study how many example utterances are necessary to correctly learn the meanings in Figure 2.

Methods

Because naturalistic data consisting of quantifiers used by parents in the presence of sets of objects is not available, we constructed simulated data by creating sets at random and sampling adult meanings according to the likelihood process described above, with $\alpha_p = \alpha_t = 0.9$. This means that the data is fairly noisy, with roughly 10% of the utterances not satisfying presuppositions and of those that do, 10% are false. We generated sets at random, each containing between 1 and 8 objects. Each object in a set was one of three animals (mouse, pig, rabbit) that was one of three colors (white, brown, or pink). Here,

the argument A was an animal and B was a color. For each set, we sampled utterances according to the target grammar: for instance, for a set containing a pink mouse and two brown rabbits, we might sample the quantifier "some" in the context " ____ mouse is pink."

For the implemented model, the prior $P(m)$ is defined by converting the grammar in Figure 4 to a *probabilistic* context-free grammar (PCFG). We assume that each non-terminal is equally likely to expand by any of its rules, except the rules that generate A and B, the arguments to the function, are 10 times as likely than other rules. This probabilistic grammar induces a probability distribution on expressions, assigning a probability to an expression corresponding to how likely it is to sample the rules required to generate the expression. This assigns short expressions higher prior probability, corresponding to the intuition that simple (concise) representations should be preferred a-priori by rational learners (Feldman, 2000; Chater & Vitányi, 2003; Goodman et al., 2008; Piantadosi et al., 2011). This prior is similar to more sophisticated versions of rule-length priors developed in other work (Goodman et al., 2008). The up-weighting of rules that generate A and B is necessary to ensure that the grammar does not generate infinitely long expressions and also to bias the learner to preferentially use the sets that are arguments to the function.

The model as described is a high-level *computational* theory of quantifier learning, not an algorithmic one (Marr, 1982). We combine several algorithmic techniques from probabilistic modeling to implement a working version of this model. This provides us with learning curves as the amount of data for the model varies, which represent the learning curves for an idealized statistical learner, operating over the space of meanings we describe. In principle, learners should be able to consider any hypothesis generated by the grammar in Figure 4. In practice, most of the hypotheses this grammar generates are very low probability, either by being long (small prior) or not explaining the data (low in the likelihood). The first approximation that we make is that our algorithm looks only at hypotheses that use 10 or fewer rule expansions[13]. We enumerate this space of hypotheses and, for computational tractability, collapse equivalent hypotheses. Thus, for instance, $\lambda\ A\ B\ .\ (\textit{doubleton?}\ A)$ is not treated distinctly from $\lambda\ A\ B\ .\ (\textit{singleton?}\ (\textit{union}\ A\ A))$. This results in 79682 hypotheses (or equivalence classes of hypotheses) that represent distinct functions on sets. This space was treated as a fixed, finite hypothesis space of expressions for purposes of inference. We note that this still represents a huge effective hypothesis space for the learner since the number of possible ways of assigning 12 word meanings to 79682 hypotheses is $79682^{12} \approx 10^{58}$.

To search through lexicons, we first ran Gibbs sampling (Geman & Geman, 1984) for varying amounts of data from 0 to 1500 sets. For each set size we ran 1000 separate Gibbs sampling runs, storing 25 lexicons with highest posterior for each run at each amount of data. This finite space of lexicons was treated as the finite hypothesis space for constructing the learning curves and results here. Note that this method means that the target lexicon had to be found at some amount of data by Gibbs sampling. However, once it is found by one run, it will be included in the final finite hypothesis space of lexicons, allowing for better statistical estimation. This is a form of selective model averaging (Madigan & Raftery, 1994), that we have used in other similar learning models (Piantadosi et al., 2011). This technique amounts to using sampling techniques to search for high probability hypotheses

[13]Hypotheses which are excluded this way have a prior probability less than 1 in 10 million.

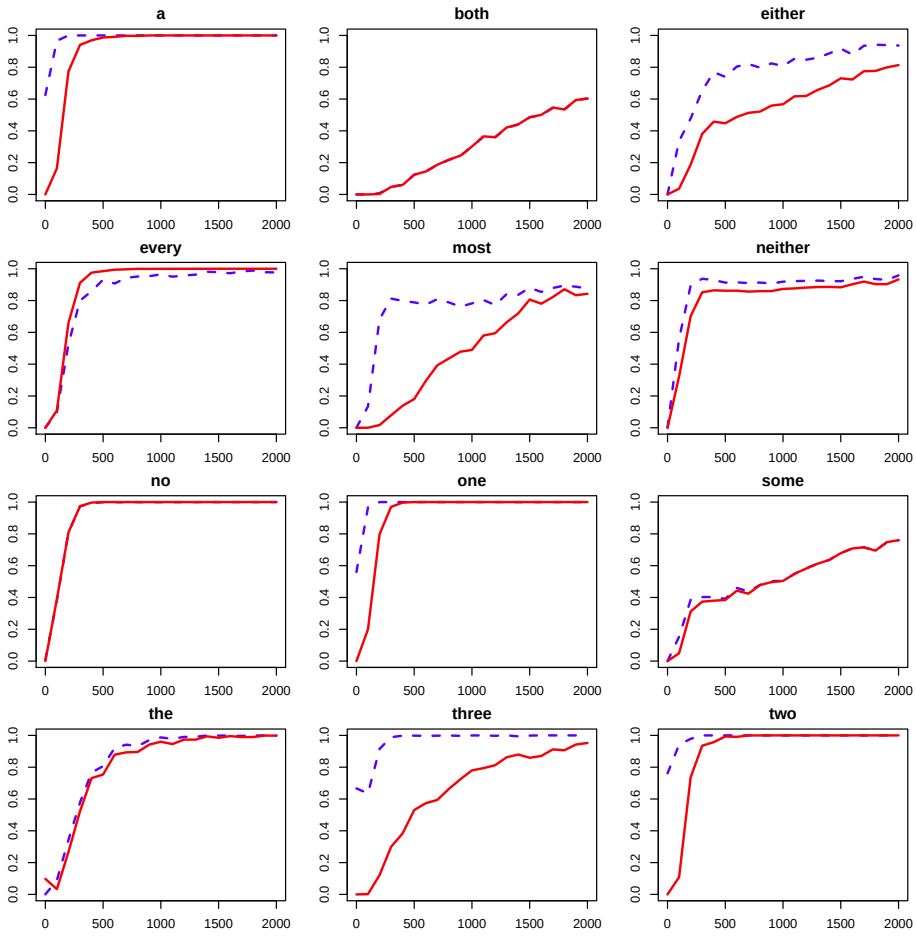

Figure 5. Learning curves for $\alpha_p = \alpha_t = 0.9$, showing model proportion correct (y-axis) versus amount of data (x-axis) for each aspect of meaning, literal (solid red) and presuppositional (dotted blue).

(at each amount of data) and then using the high probability hypotheses as a finite space for performing exact statistical inference.

Idealized learnability of quantifiers

Figure 5 shows learning curves for each of the words, on simulated data. For varying amounts of data (x-axis), these plots show the model's probability of learning the correct representation for literal meaning (red solid) and presupposition (dotted blue). The x-axis in this plot represents not the number of times that each of these words is heard, but the total number of labeled sets, only some of which are labeled with the corresponding word.

This figure shows that all of the words are learned at least 50% of the time for ideal learners who observe 2000 labeled sets. Note that the growth trend of words not near ceiling after 2000 data points indicates that with 1000 to 2000 more, they will be near ceiling. To put this total amount of data in perspective, determiners or quantifiers are used in Adam's section of the Brown corpus in CHILDES (MacWhinney, 2000) over 8000 times

and that corpus represents only a small subset of the data Adam heard. This quantity of data would be enough for an ideal learner to discover all aspects of meaning studied here, even assuming that only a quarter of instances have clear and known referents. Moreover, for the model, $\alpha_t = \alpha_p = 0.9$, meaning that only only $0.9^2 = 81\%$ of the data is perfectly "correct"—the rest is corrupted by noise. Learning through noise is possible because the model aggregates evidence from multiple contexts. Doing so is likely very important for function word learning (potentially in contrast to lexical item learning) since most function word meanings are never unambiguously conveyed in a single context.

In principle, we would like to be able to compare the model's acquisition patterns to children's, which have been studied experimentally to some degree (Karmiloff-Smith, 1981; Hanlon, 1987, 1988; Barner, Chow, & Yang, 2009; Geurts, Katsos, Cummins, Moons, & Noordman, 2010). Indeed, the broad factors which influence our model have already been hypothesized in developmental studies to affect children's learning. In one of the primary developmental studies on quantifiers, Hanlon (1987) argues that the time-frame of quantifier learning can be understood by considering general principles of cognitive development. She argues, for instance, that semantic complexity is a major determinant of acquisition, as is generality, with more specific terms acquired earlier. Both of these are also predictions of our model. In our case, semantic complexity is captured by the representation language and the model is—like a good rational learner—biased to prefer simpler hypotheses[14]. Similarly, the generality predictions can be explained by the model: all else being equal, the model will learn more specific meanings more quickly because there will be fewer competitor meanings to weed out. Hanlon also notes that in her corpus studies, frequency of usage is a major predictor of acquisition: word frequency is rank-order correlated at 0.77 with acquisition order. These frequency findings are not surprising from the perspective of a rational statistical learner, which are almost unavoidably sensitive to frequency.

In general, the model's patterns of acquisition are determined by the interaction of several factors, including the input words and frequencies (the data), the compositional representation system (the prior), and the pragmatics model (the likelihood). If all of these components were independently known, it would be possible to make detailed predictions about learning, including the relative acquisition rates for words and common acquisition errors. Unfortunately, the precise form of these components are not known: we do not know, for instance, how often children hear quantifiers in the context of different set sizes (e.g. Figure 3), or exactly what pragmatic inferences children are capable of. The learning trajectory exhibited by the model is sensitive to these components; for instance, an earlier version of this model (Piantadosi, 2011) exhibited a different learning trajectory and predicted some of children's mistakes in learning "the" (Maratsos, 1974, 1976; Warden, 1974, 1976; Karmiloff-Smith, 1981; Modyanova & Wexler, 2007; Ko, Ionin, & Wexler, 2006; Ko, Perovic, Ionin, & Wexler, 2008) and "every" (Philip, 1991, 1992, 1995; Takahashi, 1991; Philip, 1995, 1998; Kang, 1999; Philip, 2003; Fiorin, 2010; Seymour, Roeper, & De Villiers, 2003; Roeper, Strauss, & Pearson, 2004). That model used a somewhat different pragmatic setup, where word "weights" were simply memorized, instead of the Gricean setup used here. That model predicted different learning curves in part because its pragmatic assumptions

[14]We note, however, that the details of our complexity measure differs from Hanlon's, in that she argues a determinant of complexity is whether or not the reference set of a quantifier is identical to the presuppositional set. This could be incorporated into our type of framework with alterations to the grammar.

led to a different distribution of input frequencies—ones which could be fit to English word frequencies. However, we believe that the current pragmatics model is more plausible and principled, although in its current form it predicts different detailed patterns of acquisition. We therefore do not take the contribution of this work to be in making specific behavioral predictions, since the details depend on unknown factors. Instead, we see our model as providing a framework for understanding how learning with linguistically plausible semantic representations might be possible at all. As the details of children's representations, input data, and pragmatic inferences are clarified by future work, our framework will be able to produce more detailed behavioral predictions.

Constraints on quantifier meanings

Our unrestricted grammar for quantifier meanings differs from many contemporary semantic theories, which posit that potential quantifier meanings are inherently constrained by human cognitive and linguistic systems. One primary hypothesized constraint is *conservativity* (Keenan & Stavi, 1986; Barwise & Cooper, 1981): in our notation, a quantifier is conservative if it depends only on the elements of A, the first argument to the function. Thus, "Most men are happy" can be checked by looking only at the set of men[15]. Keenan and Stavi (1986) argue that conservativity provides a useful constraint for language learners. In a simple example involving sets of two individuals, they count $65,536$ possible quantifiers, only 512 of which are conservative. Intuitively, learners should benefit by narrowing down the space of possible meanings by a factor of 128. On the other hand, it may be the case that most of the quantifiers that are ruled out with a conservativity constraint are already low-probability, or that a factor of 128—7 bits of information—is not overwhelmingly useful.

Figure 6 shows a model-based analysis of how much conservativity helps under the assumptions of the idealized learning model. The black line shows correct learning (literal and presupposition together) in the unrestricted model and the blue dashed line shows a learner who only considers conservative hypotheses[16]. This plot shows that on average conservativity is not a useful constraint for this ideal learner—acquisition speed is essentially the same for the constrained and unconstrained models. We note that mathematically a smaller hypothesis space *must* help a learner since there are fewer hypotheses to consider; this shows, however, that the *degree* of help for conservativity is minimal. While there is behavioral evidence that children prefer to learn conservative quantifiers (Hunter & Conroy, 2009), this bias should not be posited for reasons of learnability (Keenan & Stavi, 1986).

This raises the question of whether any constraints on quantifier meanings could substantially aid learning. Figure 6 also shows a red dotted line, corresponding to what might be considered the most constrained learner possible—one who only considered expressions necessary for the literal meanings or presuppositions in Figure 2 as possible hypotheses. Such a learner might innately have a small collection of possible word meanings and their main challenge would be determining which word has which meaning—which of the meanings in Figure 2 meaning does "the" map to? Thus, this is the *minimal* amount of learning

[15]Conservativity is perhaps best understood by a potential counter-example to it, "only." "Only men are happy" depends on the set of things which are *not* men, violating conservativity. However, "only" is often argued not to be a quantifier due to the fact that it patterns differently in some syntactic constructions.

[16]For computational tractability, conservativity was evaluated "empirically" by evaluating the quantifier on a large collection of sets.

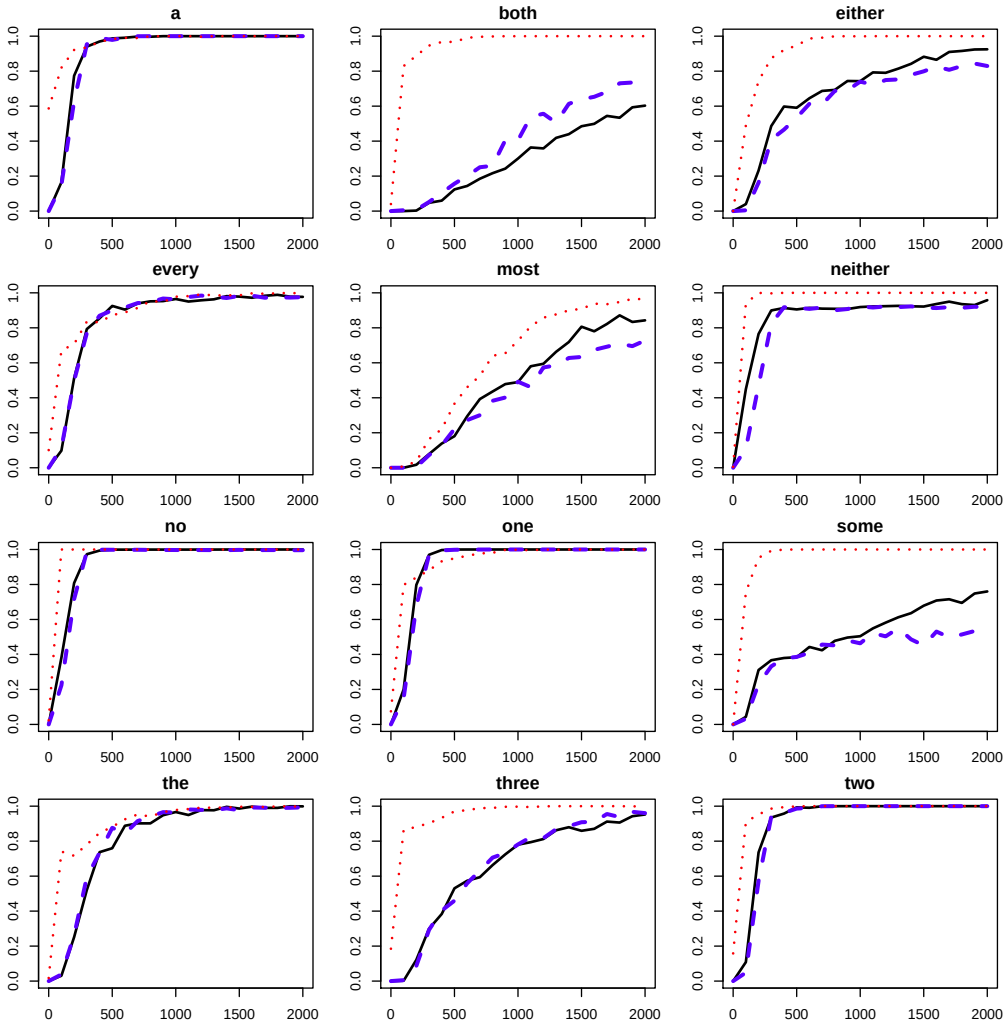

Figure 6. Learning curves for the basic unrestricted model (black), conservative quantifiers (dashed, blue), and the maximally restricted model (dotted, red). The y-axis shows probability of correct acquisition of all aspects of meanings (literal, presupposition, production probability).

that must occur, even under the most extreme nativist theories. The red dotted line in Figure 6 shows that this constrained space substantially aids the more difficult quantifiers like "both" and "some," but does comparatively little for quantifiers which are quickly learned. This indicates that in many cases, learning in the unrestricted space is not much harder than learning in the maximally restricted space. Likely, the unrestricted space has many hypotheses which are so implausible, they can be ignored quickly and do not affect learning. The hard part of learning, may be choosing between the plausible competitor meanings, not in weeding out a large space of potential meanings. An alternative way to formulate these findings is that even a maximally nativist learner who already had all and only the correct quantifier concepts, still faces a difficult acquisition problem of determining which words map to which meanings—a challenge that is on par with learning in the full unrestricted

space. Not much is gained by positing strongly innate concepts, but the capacity to learn in either system is gained by positing children have sophisticated inferential mechanisms.

Discussion

An idealized statistical learner is capable of discovering multiple unobserved aspects of semantic meaning from noisy positive evidence—word utterances in the context of sets. Our approach contrasts with previous learnability studies on quantifiers, which either require positive and negative evidence, or only provably work for a subset of meanings. We have instead approach the problem by first formalizing a space of meanings which is linguistically plausible, and then we constructed an idealized statistical learning model over that space. The learning model elegantly solves the subset problem, deals with the fact that many words have highly overlapping meanings, and handles imperfect data. Our hope is that the specific implementation we present here is not taken as a single, unitary model of quantifier acquisition. More detailed data about children's input and stages of knowledge will allow different particular implementations to be compared and empirically evaluated. The present work is a step towards developing a general framework for understanding how words like quantifiers might be acquired—by combining rich capacities for conceptual structure with genuine inductive learning.

The conceptual foundations of this approach draw heavily on work by Chater and Vitányi (2007). They proved that language learning can succeed in principle from unrestricted hypothesis spaces (consisting of all computable functions), contrasting with Gold-style (Gold, 1967) analyses. The assumptions of their approach differ from Gold's in several ways: rather than a worst-case analysis, Chater and Vitányi (2007) present average-case learnability results, and their setup characterizes the goal of the learner to be modeling the data instead of identification in the limit. Their proof works by assuming that learners attempt to find what are essentially short bit strings describing computer programs to explain the data they see. We take their results as providing the definitive theoretical argument that language is learnable from an unrestricted hypothesis space *in principle*, and their key ideas in their work have been used in behavioral predictions (see Hsu & Chater, 2010; Hsu et al., 2011). Our work complements Chater and Vitányi (2007) by constructing an example of how learning might proceed in a highly-studied linguistic domain of quantifier semantics. In contrast to their more idealized analyses, we use previously-hypothesized linguistic representations—functions that manipulate sets. Our results focus on learning such unobserved *semantic structures*, addressing what we see as one of the primary acquisition puzzles of how children learn productive linguistic constructs that do not have real-world referents. Unlike (Chater & Vitányi, 2007), we cache out computation in explicitly cognitive terms: we suppose that semantic structures are expressed in a cognitive representation language, and that the task of learners is to induce the right representation in this language for explaining adults' usage of these words. In this framework, learning is essentially a problem of *program induction*—discovering the unobserved computational process that creates the observed data (Koza, 1992)—but program induction over cognitively-plausible primitives. Our results show that this type of acquisition from a compositional space of primitives can "really work" in explaining acquisition, and we expect that compositional learning models can be generalized to learning other types of semantic and syntactic operations (for a simple example, see Piantadosi et al., 2008).

The view of "learning as program induction" is appealing because it provides an appealing compromise between nativist and empiricist theories. The model is nativist in that it "builds in" a hypothesis space of potential meanings. But hypotheses for the model need not all be explicitly represented by learners. Instead, they can be generated stochastically from the above grammar for concepts. In the best case, only two hypotheses (to be compared) must be represented simultaneously by the model—our inference algorithms have this characteristic. This amount of nativism is, in some sense, a necessity for any learning model that can arrive at the correct set of meanings—even models which do not have explicit representations build in spaces of hypotheses (see Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). However, language-specific—or, more precisely, semantics-specific—constraints and learning procedures are not *necessary* for learning. Indeed, semantic representations like quantifiers can be learned using very general principles of statistical inference like the Bayesian setup and the size principle.

As we described, the inference algorithm searches a space of around 10^{58} different possible hypotheses. A perhaps surprising fact about statistical learning algorithms is that a model which can be run in a few hours on a desktop computer, observing only a few thousand data points, can pinpoint a single hypothesis in this massive space. This is possible because the inference algorithm we used—Markov Chain Monte Carlo—stochastically "climbs" towards high probability regions of this space. So even though there are so many logically possible hypotheses for learners, nearly all of them have close to zero probability and so need not be explicitly represented. Many aren't explicitly considered by the learning algorithm because they "share" parts of other bad hypotheses—for instance, once the learning algorithm determines that "the" isn't likely to mean $\lambda\ A\ B.$ (*tripleton* (*set-difference* $A\ B$)), it needn't search through the $79682^{11} = 10^{53}$ lexica with this meaning. We suspect that something similar is true for real child language acquisition: there are many possible grammars or semantic representations that learners could consider, but most of them are very low probability and are effectively "ignored" by learning mechanisms unless the data suggests that they should be considered. Our results illustrate that idealized statistical learners can be extremely powerful in resolving a basic mystery of language learning—how it is possible at all.

One interesting feature of the learning model is that it appears in many cases to be very *easy* for the model to discover the word meanings—typically within hundreds to a few thousand labeled sets. In some cases, words that are very hard for children like "the" are not substantially delayed for the model. We have made several simplifications that likely lead to the model's comparatively rapid acquisition. First, in determining each meaning $\lambda\ A\ B\ .\ (\dots)$, we have assumed that the sets A and B are known. If learners are simultaneously learning nouns, then this would introduce more uncertainty, delaying acquisition. Second, we have assumed that the syntactic and semantic compositionality of quantifiers is known—that is, that they take two sets and assert something about their relationship. It is possible that if this also must be learned simultaneously, it would substantially complicate learning and slow acquisition. The model in its present form also has perfect memory of previous data. This assumption is not necessary for this type of LOT statistical model, but also leads to increased rates of acquisition. Finally, we have assumed relatively high rates for α_p and α_t, which determine how rapidly the model learns. Because the values of these parameters are not determined independently, the learning *rate* of the model is essentially

a free parameter.

Indeed, the ease of learning for this idealized model may raise one interesting possibility for language acquisition. It may be the case that the key puzzle for language acquisition is not the *poverty* of the stimulus, but the *abundance* of stimulus: why do some aspects of language acquisition take so long, given that an idealized statistical learner would find them so easy? Similarly, abstract syntactic principles may be learnable from surprisingly little data (Perfors, Tenenbaum, & Regier, 2011). One answer is that children are non-ideal in all sorts of ways, including memory limitations and imperfect observations. But it might be the case that even given these, facts, idealized learners find it easier than children; in which case maturational considerations—of language or other cognitive systems—might play a role. Indeed, the abundance of the stimulus was argued by Babyonyshev, Ganger, Pesetsky, and Wexler (2001) to support a maturational account of other syntactic phenomena, such as A-chain formation, since children are substantially delayed with A-chains despite their prevalence in the input. Addressing the abundance of stimulus problem is an interesting challenge for statistical learning models—one that is the polar opposite of traditional poverty arguments put forth against statistical learning. The abundance of stimuli paints a different picture of acquisition, one where the environment is full of information sources, but perhaps the hard part of language learning is using those information sources effectively.

Conclusion

This paper has studied learning problems in semantics as a case study of the logical problem of language acquisition, specifically as it applies to function words. We have argued that learners of quantifier meanings face many of the complexities that make learning language daunting: non-obvious literal meanings, the subset problem, presuppositional content, and variable word frequencies. The learning model we present posits that learners have access to a compositional system for generating possible hypothesized semantic representations, and that they use at least approximately optimal Bayesian inference to decide between those hypotheses. This provably solves the subset problem and our implementation shows that it can learn these word meanings from an unrestricted hypotheses space, in a developmentally-plausible amount of data. These general techniques could be applied to other subset-problems in language, or areas where unseen abstract structure must be inferred.

Acknowledgments

We thank Irene Heim, Ted Gibson, Melissa Kline, Avril Kenney, and Celeste Kidd for helpful discussions during various stages of this work, and feedback on earlier drafts of this paper. This research was supported by a National Science Foundation graduate research fellowship and a Social, Behavioral & Economic Sciences Doctoral Dissertation Research Improvement Grant in linguistics (to S.T.P.).

References

Angluin, D. (1987). Learning regular sets from queries and counterexamples. *Information and computation*, *75*(2), 87–106.

Babyonyshev, M., Ganger, J., Pesetsky, D., & Wexler, K. (2001). The maturation of grammatical principles: Evidence from Russian unaccusatives. *Linguistic Inquiry*, *32*(1), 1–44.

Barner, D., Chow, K., & Yang, S. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive psychology*, *58*(2), 195–219.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and philosophy*, *4*(2), 159–219.

Bergen, L., Goodman, N., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the thirty-fourth annual conference of the cognitive science society.*

Berwick, R. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

Blackburn, P., & Bos, J. (2005). *Representation and inference for natural language: A first course in computational semantics*. Center for the Study of Language and Information.

Borer, H., & Wexler, K. (1987). The maturation of syntax. In T. Roper & E. Williams (Eds.), *Parameter setting* (Vol. 4, p. 123). Norwell, MA: Kiuwer Academic Publishers.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.

Chater, N., & Vitányi, P. (2007). Ideal learning'of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*(3), 135–163.

Clark, R. (1996). Learning first order quantifier denotations: An essay in semantic learnability. *IRCS Technical Report 96-19*.

Clark, R. (2010). *Some computational properties of generalized quantifiers.* (Unpublished manuscript)

Cover, T., & Thomas, J. (2006). *Elements of information theory*. Hoboken, NJ: John Wiley and sons.

Crain, S. (1992). The semantic subset principle in the acquisition of quantification. In *Workshop on the Acquisition of WH-Extraction and Related Work on Quantification, University of Massachusetts, Amherst, MA*.

Crain, S. (1993). Semantic subsetutions. In *Invited paper presented at the Center for Cognitive Science Conference: Early Cognition and the Transition to Language, University of Texas, Austin.*

Crain, S., Ni, W., & Conway, L. (1994). Learning, parsing and modularity. *Perspectives on sentence processing*, 443–467.

Crain, S., & Philip, W. (1993). Global semantic dependencies in child language. In *GLOW Colloquium* (Vol. 16). Lund, Sweden.

Crain, S., & Thornton, R. (2000). *Investigations in Universal Grammar: A guide to experiments on the acquisition of syntax and semantics*. Cambridge, MA: MIT Press.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Fiorin, G. (2010). Meaning and dyslexia: a study on pronouns, aspect, and quantification. *Lot Dissertation Series*, *235*.

Florêncio, C. (2002). Learning generalized quantifiers. In *Proceedings of Seventh ESSLLI Student Session.*

Frank, M., & Goodman, N. (2012). Quantifying pragmatic inference in language games. *Science*, *336*.

Frank, M., Goodman, N., & Tenenbaum, J. (2007). A Bayesian framework for cross-situational word learning. *Advances in neural information processing systems*, *20*.

Geman, S., & Geman, D. (1984). Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(2), 721–741.

Geurts, B., Katsos, N., Cummins, C., Moons, J., & Noordman, L. (2010). Scalar quantifiers: Logic, acquisition, and processing. *Language and cognitive processes*, *25*(1), 130–148.

Gierasimczuk, N. (2007). The problem of learning the semantics of quantifiers. *Logic, Language, and Computation*, 117–126.

Gold, E. (1967). Language identification in the limit. *Information and control*, *10*(5), 447–474.

Goodman, N., & Stuhlmüller, A. (2012). Knowledge and implicature: Modeling language understanding as social cognition. In *Proceedings of the thirty-fourth annual conference of the cognitive science society.*

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108–154.

Goodman, N., Ullman, T., & Tenenbaum, J. (2009). Learning a theory of causality. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2188–2193).

Grice, H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics, Vol. 3, Speech Acts* (pp. 41–58). New York: Academic Press.

Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci*, *14*(10.1016).

Gualmini, A., & Schwarz, B. (2009). Solving learnability problems in the acquisition of semantics. *Journal of Semantics*.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics*, *17*(1), 63–98.

Hale, M., & Reiss, C. (2003). The Subset Principle in phonology: why the tabula can't be rasa. *Journal of Linguistics*, *39*(02), 219–244.

Hanlon, C. (1987). Acquisition of set-relational quantifiers in early childhood. *Genetic, social, and general psychology monographs*.

Hanlon, C. (1988). The emergence of set-relational quantifiers in early childhood. *The development of language and language researchers: Essays in honor of Roger Brown*, 65–78.

Heim, I. (1991). Artikel und definitheit. *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*, 487–535.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Malden, MA: Wiley-Blackwell.

Horning, J. (1969). *A study of grammatical inference*.

Hsu, A., & Chater, N. (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive science*, *34*(6), 972–1016.

Hsu, A., Chater, N., & Vitányi, P. (2011). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition*, *120*(3), 380–390.

Hunter, T., & Conroy, A. (2009). Children's restrictions on the meanings of novel determiners: An investigation of conservativity. In *BUCLD* (Vol. 33, pp. 245–255).

Jeffreys, S. (1961). *Theory of probability*. Oxford: Oxford University Press.

Kang, H. (1999). Quantifier spreading by English and Korean children. *Ms., University College, London*.

Karmiloff-Smith, A. (1981). *A functional approach to child language: A study of determiners and reference* (Vol. 24). Cambridge, UK: Cambridge University Press.

Karttunen, L., & Peters, S. (1979). Conventional implicature. *Syntax and semantics*, *11*, 1–56.

Katz, Y., Goodman, N., Kersting, K., Kemp, C., & Tenenbaum, J. (2008). Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of Thirtieth Annual Meeting of the Cognitive Science Society.*

Keenan, E., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and philosophy*, *9*(3), 253–326.

Keenan, E., & Westerståhl, D. (1997). Generalized quantifiers in linguistics and logic. *Handbook of logic and language*, 837–893.

Kemp, C., Goodman, N., & Tenenbaum, J. (2008). Learning and using relational theories. *Advances in neural information processing systems*, *20*, 753–760.

Ko, H., Ionin, T., & Wexler, K. (2006). Adult L2-learners lack the maximality presupposition, too. In

Proceedings of the Inaugural Conference on Generative Approaches to Language Acquisition–North America, Honolulu, HI (pp. 171–182).

Ko, H., Perovic, A., Ionin, T., & Wexler, K. (2008). Semantic universals and variation in L2 article choice. In *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference (GASLA 9)* (pp. 118–129).

Koza, J. (1992). *Genetic programming: on the programming of computers by means of natural selection*. Cambridge, MA: MIT Press.

Kwiatkowski, T., Goldwater, S., & Steedman, M. (2009). Computational Grammar Acquisition from CHILDES data using a Probabilistic Parsing Model. In *Psychocomputational Models of Human Language Acquisition (PsychoCompLA)*.

Langford, J., & Holmes, V. (1979). Syntactic presupposition in sentence comprehension. *Cognition*, *7*(4), 363–383.

Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications*. New York: Springer-Verlag.

Ludlow, P., & Neale, S. (2008). Descriptions. *The Blackwell Guide to the Philosophy of Language*, 288–313.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Hillsdale, New Jersey.

Madigan, D., & Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.*, *89*, 1535–1546.

Maratsos, M. (1974). Preschool children's use of definite and indefinite articles. *Child Development*, *45*(2), 446–455.

Maratsos, M. (1976). *The use of definite and indefinite reference in young children: An experimental study of semantic acquisition*. Cambridge, UK: Cambridge University Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman & Company.

Modyanova, N., & Wexler, K. (2007). Semantic and pragmatic language development: Children know 'that' better. In *Proceedings of the 2nd conference on generative approaches to language acquisition–north america (galana 2)* (pp. 297–308).

Montague, R. (1973). The Proper Treatment of Quantification in Ordinary English. *Formal Semantics*, 17–34.

Mostowski, A. (1957). On a generalization of quantifiers. *Fundamenta mathematicae*, *44*(1), 12–36.

Mostowski, M. (1998). Computational semantics for monadic quantifiers. *Journal of Applied Nonclassical Logics*, *8*, 107–122.

Musolino, J. (2006). On the semantics of the Subset Principle. *Language Learning and Development*, *2*(3), 195–218.

Musolino, J., Crain, S., & Thornton, R. (2000). Navigating negative quantificational space. *Linguistics*, *38*(1), 1–32.

Perfors, A., Tenenbaum, J., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*.

Philip, W. (1991). Spreading in the Acquisition of Universal Quantifiers. *West Coast Conference on Formal Linguistics*, *10*, 359–373.

Philip, W. (1992). Distributivity in the Acquisition of Universal Quantifiers. *Proceedings of the 2nd Conference on Semantic and Linguistic Theory, Ohio State Working Papers in Linguistics*, *40*, 327–346.

Philip, W. (1995). *Event Quantification in the Acquisition of Universal Quantification*. Ph.D. thesis, University of Massachusetts, Amherst.

Philip, W. (1998). The wide scope interpretation of postverbal quantifier subjects: QR in the early grammar of Spanish. *Proceedings of GALA 1997*.

Philip, W. (2003). Specific indefinites & quantifier scope for children acquiring Dutch and Chinese.

In *Meeting of the Linguistic Society of the Netherlands. In V. van Geenhoven (ed.), Semantics MeetsAcquisition.* Dordrecht, The Netherlands: Kluwer.

Piantadosi, S. (2011). *Learning and the language of thought.* Unpublished doctoral dissertation, MIT.

Piantadosi, S., Goodman, N., Ellis, B., & Tenenbaum, J. (2008). A Bayesian model of the acquisition of compositional semantics. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*

Piantadosi, S., Tenenbaum, J., & Goodman, N. (2009). Beyond Boolean Logic: Exploring Representation Languages for Learning Complex Concepts. In *Proceedings of the thirtieth annual conference of the cognitive science society.*

Piantadosi, S., Tenenbaum, J., & Goodman, N. (2011). Bootstrapping in a language of thought: a formal model of numerical concept learning. *Cognition*, *123*, 199–217. Available from `http://web.mit.edu/piantado/www/papers/piantadosi2012bootstrapping.pdf`

Piantadosi, S., Tenenbaum, J., & Goodman, N. (2012). Bootstrapping in a language of thought: a formal model of numerical concept learning. *Cognition*, *123*, 199–217. Available from `http://web.mit.edu/piantado/www/papers/piantadosi2012bootstrapping.pdf`

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The Meaning of 'Most': Semantics, Numerosity and Psychology. *Mind & Language*, *24*(5), 554–585.

Roeper, T., Strauss, U., & Pearson, B. (2004). The acquisition path of quantifiers: Two kinds of spreading. *Current Issues in Language Acquisition, UMOP*, *34*.

Russell, B. (1905). On denoting. *Mind*, *14*(56), 479–493.

Russell, B. (1957). Mr. Strawson on referring. *Mind*, *66*(263), 385.

Sauerland, U. (2003). Implicated presuppositions. In *Proceedings of the conference on Polarity, Scalar Phenomena, Implicatures. University of Milano Bicocca.*

Schlenker, P. (2006). Maximize presupposition and Gricean reasoning. *Manuscript, UCLA and Institute Jean-Nicod, Paris*.

Seymour, H., Roeper, T., & De Villiers, J. (2003). *Diagnostic evaluation of language variation.* Thieme Medical Publishers, NY.

Singh, R. (2009). Maximize Presupposition! and local contexts. *Natural Language Semantics*, 1–20.

Siskind, J. (1996). A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings. *Cognition*, *61*, 31–91.

Smith, K., Smith, A., Blythe, R., & Vogt, P. (2006). Cross-situational learning: a mathematical approach. *Symbol grounding and beyond*, 31–44.

Smolensky, P. (1996). *The initial state and 'richness of the base' in Optimality Theory* (Tech. Rep.). Johns Hopkins University, Department of Cognitive Science. JHU-CogSci-96-4.

Steedman, M. (2000). *The syntactic process* (Vol. 131). Cambridge, MA: MIT Press.

Strawson, P. (1950). On referring. *Mind*, *59*(235), 320–344.

Takahashi, M. (1991). *Children's interpretation of sentences containing every.* Amherst, MA: GLSA.

Tenenbaum, J. (1999). *A Bayesian Framework for Concept Learning.* Ph.D. thesis, Massachusetts Institute of Technology.

Tiede, H. (1999). Identifiability in the limit of context-free generalized quantifiers. *Journal of Language and Computation*, *1*(1), 93–102.

Ullman, T., Goodman, N., & Tenenbaum, J. (2010). Theory Acquisition as Stochastic Search. In *Proceedings of thirty second annual meeting of the cognitive science society.*

van Benthem, J. (1984). Semantic automata. In J. Groenendijk, D. d. Jongh, & M. Stokhof (Eds.), *Studies in discourse representation theory and the theory of generalized quantifiers.* Dordrecht, The Netherlands: Foris Publications Holland.

van Benthem, J. (1986). *Essays in logical semantics.* Dordrecht, The Netherlands: Reidel.

Vogt, P., & Smith, A. (2005). Learning colour words is slow: A cross-situational learning account. *Behavioral and Brain Sciences*, *28*(04), 509–510.

Von Fintel, K. (2004). Would you believe it? The king of France is back! Presuppositions and truth-value intuitions. *Descriptions and beyond*, 315–341.

Warden, D. (1974). *An experimental investigation into the child's developing use of definite and indefinite referential speech.* Ph.D. thesis, University of London.

Warden, D. (1976). The influence of context on children's use of identifying expressions and references. *British Journal of Psychology.*

Wexler, K., & Manzini, M. (1987). Parameters and learnability in binding theory. *Parameter setting*, 41–76.

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

Yu, C., & Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.

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

Zettlemoyer, L. S., & Collins, M. (2005). Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. In *UAI* (pp. 658–666).