Bayesian Inference for Bilingual Word Learning

Sebastian Rolotti, McNair Scholar
The Pennsylvania State University

McNair Faculty Research Advisor:
Ping Li, Ph.D
Professor of Psychology, Linguistics, and Information Sciences and Technology
Department of Psychology
College of Liberal Arts
The Pennsylvania State University

Abstract

The indeterminacy problem describes the challenge that infants face in deciding which words refer to which objects in their environment. Bayesian models use probabilistic inferences to resolve this induction problem and show improved performance over other computational models in constructing potential lexicons and inferring speakers' referential intentions. In this study we investigate a Bayesian model's ability to learn in more complex situations, first with more objects than in previous research and then in a bilingual scenario where more than one word refers to the same object. We found that the model's absolute and relative performance was attenuated with increased complexity.

Introduction

Communication through language, even between two native speakers, can often be difficult and opaque. In an attempt to illustrate this notion, Wittgenstein says this, "Language disguises thought. So much so, that from the outward form of the clothing it is impossible to infer the form of the thought beneath it" (1922, pg. 22). For an infant with no prior access to language or understanding of social behavior, determining the intentions of speakers and the words that correspond to those intentions is doubly difficult. In his famous formulation of the problem, Quine (1960) imagines being with a foreign language speaker who points and says *gavagai* upon coming across a rabbit. Quine's indeterminacy problem, faced by infant word learners, is the problem of figuring out just what *gavagai* refers to – the rabbit itself, the rabbit's tail or ear, the whiteness of its fur, or perhaps mammals or animals in general. Of these infinite possible referents, what strategies does a word learner make use of to pick out the right one? The challenge of word-to-world mapping becomes more difficult when one considers the small proportion of words an infant hears that actually refer to objects in his or her immediate surroundings, and how irregularly those few words co-occur with those objects (Yu, 2008). With so much information to extract structure from, both Quine and Yu point out that an infant's learning system must be advantageously constrained in some way.

Constraints in Monolingual Word Learning

Despite controversies that surrounds their origins and the degree to which they are actually used, researchers for the most part agree on the need to posit a number of important assumptions that help word learners constrain a potentially infinite problem space (Markman, 1994). The *whole object constraint* assumes that words refer to whole objects instead of parts of objects. The *taxonomic constraint* assumes that novel words might be generalizable to objects that are similar to each other. The *mutual exclusivity constraint* assumes that each object has only one label (Markman, 1994). A number of similar constraints have also been posited, such as the *contrast constraint* (Clark, 1993) and the *novel-name nameless-category constraint*,(Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992) but these constraints explore different motivations yielding mutual exclusivity, i.e. that novel words refer to novel objects (Yu, 2008; Byers-Heinlein & Werker, 2009).

There is strong behavioral evidence to suggest that monolinguals make heavy use of the mutual exclusivity constraint when deciding upon the referents of words, with the constraint developing over time through its communicative function (Davidson & Tell, 2005). The mutual exclusivity constraint appears to be available to infant word learners from at least 16 months of age (Liittschwager & Markman, 1994). It is important to note, though, that mutual exclusivity is by no means an absolute condition for word-object mapping. Yurovsky and Yu (2008) have demonstrated adult violations of mutual exclusivity in mapping a label to two distinct objects (a case of homonymy) across situations, though the constraint was used to pick an object out among co-occurring referents within individual trials. Liittschwager and Markman (1994) characterize mutual exclusivity as a 'default assumption', showing that difficulties in learning a second label for familiar objects (i.e., synonymy) disappear with enough evidence and processing capacity. They concluded that "mutual exclusivity works as a probabilistic bias and not as an absolute constraint" (pg. 957).

Parents direct somewhere between 300 to 400 utterances an hour on average to their children (Hart & Risley, 1995). With exposure to so many words over such a small time span, an intractable level of ambiguity about intended referents is likely to persist, even with the help of a number of social, linguistic, and conceptual constraints (Smith & Yu, 2008). As a means of helping to resolve this issue, Aslin and Newport (2012) point out that children are highly proficient at extracting organizational structure from ambiguous data from mere observation. The process by which children extract this information about distributions in the input is referred to as *statistical learning* (Aslin & Newport, 2012; Saffran, Aslin, & Newport, 1996). Although their probabilistic reasoning based on this information differs from that of adults (likely due to cognitive limitations), children will naturally sample these distributional properties even without the presentation of a specific task (David, Newport, & Aslin, 2009).

Computational Models

In the effort to understand the problems that word learners face and the statistical mechanisms by which these problems may be solved, computational models have

provided the ability to systematically control and manipulate relevant variables, flexibly test a range of hypotheses, and for many problems change from slow and descriptive to experimental methods (Zinszer & Li, 2010; Li & Zhao, 2012a). Whether or not these models accurately depict the processes of the cognitive system, they allow us to understand the goals and constraints faced by the system and to compare human performance to the models' optimized reasoning (Perfors, Tenenbaum, Griffiths, & Xu, 2011). Furthermore, models allow researchers to better understand the implications of their ideas, assumptions, and simplifications and thereby elaborate the phenomena under investigation to develop further questions for behavioral and neurological research. Computational models thus have a reciprocal relationship with empirical research, being informed by earlier findings and data and informing later studies (Li & Zhao, 2012a).

Existing models of word learning can generally be delineated into two classes: *hypothesis elimination* models and *associative* models (Xu & Tenenbaum, 2007a). Hypothesis elimination models generate a number of hypotheses, completely eliminating through deductive inference those that do not fit the observed data. Siskind (1996) developed an especially in-depth treatment of models of this type, presenting a formal algorithm for keeping track of only those hypotheses that provide valid potential solutions. The second class, the associative models, learn words by tracking co-occurrence or similarity statistics across situations. These co-occurrences can either be between words and objects in the environment (Roy & Pentland, 2002) or between a word and the other words in the surrounding linguistic input (Li, Burgess, & Lund, 2000). One important and broadly-applied family of associative models are connectionist networks, which map words to objects (local representations) or a group of perceptual features (distributed representations) through a form of gradient descent (Elman, 1996; Seidenberg, 1989; Li, 2009; see Li and Zhao 2012b for review). Among other domains, connectionist models have successfully simulated several phenomena of bilingual word learning, reviewed below.

As a means of evaluation, Xu & Tenenbaum (2007a) introduce five core word-learning phenomena that must be replicated by any valid computational model: (1) learning inductively from very few examples, (2) learning from only positive examples, i.e., they are never told what words do *not* refer to, (3) learning a system of overlapping concepts, (4) learning word meanings in a graded fashion, with varying degrees of confidence depending on the number and quality of examples available, and (5) learning based on intentional reasoning and how the examples are being generated. Xu and Tenenbaum go on to evaluate each class of models for each of these criteria, concluding that neither is capable of succeeding in all five areas. The use of connectionist models has also been critiqued because of the opacity of the networks' solutions; more generally, researchers have questioned the wisdom of bottom-up approaches such as these when so little is known about the degree to which they accurately model the physiological mechanisms of the brain (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010).

Bayesian Models

Frank, Goodman, & Tenenbaum (2009) introduce another distinction between *social intentional* theories which emphasize a rich understanding of speakers' intentions to learn

words, and *cross-situational* approaches, the foundational underpinning of most computational models, which depend largely on co-occurrence, viewing speakers' intentions as ambiguous noise to be canceled out through multiple observations. In order to bridge the gap between these two theoretical approaches, Frank et al. (2009) propose that the learning of words and intentions be combined into a single joint-inference problem, to be solved with a new class of Bayesian inference models. Models of this type have been shown to be capable of accounting for all five of the core word learning phenomena (Xu & Tenenbaum, 2007a) and in comparison tests have been shown to outperform a number of models belonging to the other two classes in developing a lexicon and figuring out speakers' intentions (Frank et al., 2009). The same study also demonstrated this model's capability to display a number of behaviorally realistic learning phenomena, such as graded mutual exclusivity and *fast mapping*, or learning an object-label association from a single observational trial.

Bayesian models are similar in principle to hypothesis elimination models, except that they evaluate hypotheses probabilistically. All hypotheses are considered, with each being assigned a *posterior probability*, indicating the model's degree of belief in that specific hypothesis (Perfors et al., 2011). Instead of eliminating any hypotheses, the model assigns smaller and smaller probabilities to the more unlikely hypotheses. Posterior probabilities across all hypotheses must sum to one, in accordance with the *principle of conservation of rational belief* (Xu & Tenenbaum, 2007a; Perfors et al., 2011), requiring that more confidence in one hypothesis is balanced by a lower degree of confidence in others.

The posterior probability $p(H_i|D)$ of a hypothesis H_i given data D, can be calculated according to Bayes' Rule:

$$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{\sum_j p(D|H_j)p(H_j)}$$

Here $p(D|H_i)$ is known as the *likelihood probability*, while $p(H_i)$ is known as the *prior probability*. The likelihood probability captures how much one would expect to observe the data D if H_i were true, while the prior captures how likely hypothesis H_i is before observing any data at all. The hypothesis space from which the hypotheses and their associated prior probabilities are drawn from is structured in accordance with the modelers' assumptions about the word learning principles and constraints available to the model (Xu & Tenenbaum 2007a). For example, hypotheses with smaller lexicons or lexicons that do not give two labels to the same object in accordance with the mutual exclusivity constraint may be allotted a higher prior probability. The product of the likelihood and prior expresses a trade-off between how well the hypothesis, in our case a lexicon, fits the observed data on the one hand and how inherently complex the lexicon is on the other (Perfors et al., 2010). The better the fit and the simpler the hypothesis, the higher the probability that it is the right one. The denominator in the above expression normalizes the term, ensuring that the probabilities sum to one. While this expression represents an evaluation of the hypotheses, how they are generated from a possibly infinite hypothesis space is another model-dependent question to be addressed below.

Of great importance in Bayesian inference are the related notions of the *size principle* and *suspicious coincidence*. The size principle dictates that smaller hypotheses assign greater likelihood probabilities than do larger hypotheses, and this difference becomes exponentially large as the data set upon which the hypotheses are based increases (Xu & Tenenbaum, 2007a). This is more clearly expressed through the intuitive notion of suspicious coincidence, whereby the model is sensitive to the way data are being generated, assuming for the most part that the objects to be labeled are random, representative samples of referents for that word. As a result of this assumption, the smallest hypothesis that fits the data becomes the most attractive. For example, after the native in Quine's (1960) example points to the rabbit and says *gavagai*, the word learner will probably assume it refers to the rabbit and not to all animals or living things, but his or her level of certainty in that assumption would be rather low. After coming across two more rabbits and hearing the same label while other animals do not receive that name, the word learner should become more certain of the correspondence between *gavagai* and rabbits. If the word did truly refer to all animals, it would be a suspicious coincidence that of all the animals that could have been randomly selected and presented in association with *gavagai*, all three would be rabbits. This type of *graded generalization* is characteristic of human word learners, providing a partial solution to the problem of no negative evidence (Perfors et al., 2011), and is an important feature of Bayesian models that distinguishes them from previous model classes (Xu & Tenenbaum, 2007b).

Behavioral Evidence for Bayesian Models

Bonawitz & Griffiths (2010) indicate that current research proposes that inductive problem solving of the type found in word learning is Bayesian in character. Xu & Tenenbaum (2007a) presented 3-4 year old children with either one or three similar objects and a single associated label, asking them to either pick out other objects that would also be named by that label or to judge whether a newly presented object would fit with that label. They found in both cases that in the three-object trial the children tended to generalize the name to refer to less similar objects significantly less often than in the one-object trial, indicating that the children were reasoning with the notion of suspicious coincidence in mind and narrowing in on the most specific valid hypothesis given more data. That is, children shown only one instance of a 'dog' may be willing to generalize the word to a cat or bear, but after observing three examples of a 'dog' children would tend to associate the word 'dog' with more specifically dog-like qualities and expect cats and bears to have a different label. Xu & Tenenbaum (2007b) further demonstrated the Bayesian nature of word learning by finding that four year olds only changed their generalizations according to suspicious coincidence in this way when they had reason to believe the sampling of objects was representative of the entire space of referents for a label, i.e., chosen by a "teacher" (*strong sampling*) instead of picked out by some other accidental, less representative process, i.e., chosen based on similarity to previously labeled referents (*weak sampling*), indicating a sensitivity to the data generation process.

A number of studies have also shown the cognitive plausibility of models of this type. Adults and 12-14 month old children have been shown capable of tracking co-occurrence statistics over a number of ambiguous trials of multiple-word to multiple-

object pairings to systematically learn an entire lexicon of word-referent mappings (Yu & Smith, 2007; Smith & Yu, 2008). These results are partly explained by a demonstrable ability, at least in adults, to use partial knowledge from preceding situations to not only better learn labels for previously presented objects, but also to constrain the name possibilities of novel objects to support the systematic learning of an entire lexicon (Yurovsky, Fricker, Yu, & Smith, 2010). The systematic nature of lexicon learning is represented in the model considered in this paper by its comparative evaluation of whole lexicons instead of single word-object mapping entries in the lexicon. Vouloumanos (2008) established the plausibility of keeping track of and considering the probability of a number of candidate word-referent mappings with a high degree of exactness, even when those mappings were extremely unlikely. A graded version of constraints such as mutual exclusivity is therefore more likely to represent human word learning than the strictly discrete all-or-nothing constraints reflected in previous hypothesis generation models.

Bilingual Word Learning

While models of monolingual word learning abound, less work has been done to model bilingual word learning (Li & Zhao, 2012a), a process which fosters the development of a different set of word learning skills, constraints, strategies, and expectations (Merriman & Kutlesic, 1993; Bialystok, Barac, Blaye, Poulin-Dubois, 2010). Davidson and Tell (2005) explain that use of the mutual exclusivity constraint might be problematic for bilinguals who, if adhering to this assumption, would be hesitant to assign another name to an object previously labeled in another language, though this would be necessary to properly learn two languages. They found that bilingual children are much less likely to depend on the mutual exclusivity constraint than monolinguals, who make use of it in nearly all cases, particularly as they get older. Byers-Heinlein and Werker (2009) suggest that the acquisition of translation equivalents for bilinguals and trilinguals precedes the development of mutual exclusivity, and that the number of these equivalents in a lexicon is likely related to the degree to which mutual exclusivity is obeyed. Au and Glusman (1990) demonstrated, though, that both monolingual and bilingual five year olds readily accepted two labels for an object when the names explicitly came from different languages, indicative of a sensitivity to sampling and a graded sense of mutual exclusivity (Xu & Tennenbaum, 2007b). While there is research to show an early language differentiation in pragmatic abilities and in the organization of the lexicon (Paradis, 2001), it remains to be seen how early and through what means monolingual and bilingual infants are able to distinguish the language origin of different words, and when and to what degree this knowledge facilitates bilingual word learning (Byers-Heinlein and Werker, 2009; Perfors, 2001). The model discussed in this paper therefore makes no assumptions regarding the infant word learner's linguistic meta-knowledge (although such knowledge could be highly important in guiding bilingual learning).

Research has also shown that while the course and rate of language development for monolingual and bilingual children are similar, the lexicon of bilingual children in each language is smaller than that of a comparable monolingual (Bialystok et. al, 2010), though the bilingual may know the same or more words when both languages are taken into account (Byers-Heinlein and Werker, 2009). This may be due to the division of the

bilingual's experience between two languages or due to a difference in the process of vocabulary learning (Bialystok et. al, 2010).

Models of Bilingual Word Learning

While some bilingual models of language acquisition and word learning do exist, these models typically have connectionist architectures and tend to be more concerned with the representational differences between monolinguals and bilinguals instead of the different constraints and trajectories of the word learning process (see Li & Zhao, 2012a for a brief review). Zhao and Li (2007) and Li (2009), for example, used a temporally dynamic approach to show inter- and intra-language competition effects in a self-organizing connectionist network and the consequences these effects, along with a number of word learner variables such as age of onset, have on a bilingual's lexical representation. Other self-organizing connectionist network models have given accounts of individual differences due to working memory and proficiency, priming and interference effects (Li & Farkas, 2002), critical development periods (Richardson & Thomas, 2008), and taxonomic responding and fast mapping in early word learning (Mayor & Plunket, 2010). Aside from the criticisms of connectionist models discussed earlier, Yu (2008) observes that many simulation studies of this type base themselves on artificial data that presuppose word-object pairings, failing to address the inductive mapping problem we are presently considering. Byers-Heinlein and Werker (2009) further note that no computational account has yet addressed mutual exclusivity in the multilingual situation.

Current study

To our knowledge, no studies have yet assessed the capability of Bayesian models to give a faithful depiction of the bilingual word learning process, one which is clearly different from and more complex than the monolingual case. As a first step, in this study we adapt existing computational models of monolingual language processing to the bilingual situation (Brysbaert, Verreyt, & Duyck, 2010), and, specifically, extend the findings of Frank et al.'s (2009) Bayesian intentional model, which is clearly documented and has been shown to be more effective at choosing lexicons than previous model classes. We further apply the model to a bilingual data set to assess the extent to which it is able to perform in this more complex case. Finally, we discuss the assumptions of the model which result in its varied performance in monolingual and bilingual contexts of varying complexity.

Methods

In the present study, we consider the model presented by Frank and colleagues (2009) under a new set of stimuli. In the first experiment, we modestly increase the complexity of the model's input, drawing parent-child interactions from those used by Fernald and Morikawa (1993), similar to the Rollins section of the Child Language Data Exchange System (CHILDES) used by Frank et al. (2009) (MacWhinney 2000). Using the same criteria employed in the original study, we evaluate the model's performance

and compare it to previous findings. In the second experiment, we introduce a second language by translating approximately 50% of the training material from Experiment 1 into Spanish. Again, we evaluate the model's performance on this new task and compare it with previous results.

While the original study and its corresponding supplementary material should be consulted for specifics of the model's design and implementation, we note a number of assumptions that are made for the sake of clarity. A further treatment of some of these assumptions is provided in the Discussion section, as they apply to the results of our simulations.

Model Assumptions

In order to map words to objects, we must first assume that the model is already capable of (1) parsing speech and (2) distinguishing objects in the first place. A number of behavioral results indicate that by 17 months of age, most typically developing children accomplish both of these feats, marking the approximate onset of the so-called vocabulary spurt. In a seminal paper on statistical learning, Saffran, Aslin, & Newport (1996) showed that eight month olds were able to parse and group three-syllable strings through an experience independent mechanism after only two minutes of exposure to an artificial language. The infants were able to do this by estimating and tracking the conditional or transitional probabilities of one syllable following another, parsing between low probability pairs (Swingley, 2009; Aslin & Newport, 2012). Extending these findings beyond artificial language, Hay, Pelucchi, Estes, & Saffran (2011) went on to show that 17 month olds were able to track bidirectional transitional probability statistics from two minutes of exposure to an unfamiliar natural language (i.e., Italian) to parse words and then later treat them as labels for novel objects. Infants have also been observed to discriminate familiar and novel sequences of shapes by two months (Kirkam, Slemmer, & Johnson, 2002).

Rosch et al. (1976) distinguish many levels of abstraction along the hierarchical object taxonomy, including, from low to high-level, subordinate, basic, and superordinate (e.g., "Tigger", cat, animal). Through a series of experiments, Rosch et al. show that basic objects share the largest number of common attributes, are the earliest categories perceived, sorted, and named by children, and are the most necessary and commonly used in language. Markman and Wachtel (1998) point out that basic level categories are commonly mutually exclusive, and that their use as a primary means of learning word-to-world mappings reasonably fits an assumption of mutual exclusivity. Generalization tendencies and a preference for the basic level are further explored by Xu & Tenenbaum (2007a), and in this model we assume that only basic objects are being considered by the word learner (Frank et al., 2009).

Lastly, it should be noted that while there are many possible considerations of word "meaning", including lexical co-occurrence (Li et al., 2000) and a more intensional account through groupings of distributed perceptual features (Li et al., 2007; Li 2009), "meaning" is here assumed to be extensional, i.e., the scope of the referents a label picks out (Xu & Tenenbaum, 2007a; Frank et al., 2009).

Levels of analysis

It is important to keep in mind that Bayesian models describe the strategies or approaches which may be applied when encountering new information, rather than making claims or commitments about the psychological or physiological mechanisms by which people actually learn and reason (Bonawitz & Griffiths, 2010; Bonwitz & Griffiths, 2010; Frank et al., 2009; Griffiths et al., 2010; McClelland, 2009; Xu & Tenenbaum, 2007a). Clark (1989) maintains that "explanation is…a matter of depicting the structure *at the right level*. And the right level here is determined by the need to capture *generalizations* about the phenomena picked out by the science in question" (pg. 181). Marr's (1982) Tri-Level Hypothesis classifies all information processing systems (the cognitive system included) into three levels of analysis: (1) computational, (2) algorithmic, and (3) implementational or physical. Computational analysis involves understanding the system's problems, goals, and motivations. Algorithmic analysis involves understanding the representations the system uses to solve those problems and how it goes about building and manipulating those representations. Implementational analysis involves understanding how the system's hardware functions, manifested as neurophysiological research in the cognitive case. Typically, Bayesian models of word learning should be taken as computational level models, or program explanations in the words of Jackson and Pettit(1988) and Clark (1989), that show what problems face the word learner and outlining the common features of general strategies for overcoming them (Xu & Tenenbaum, 2007a; Griffiths et al., 2010).

Model Design

As in Frank et. al (2009), the intentional model's parameters dictating the probability that words are used referentially and the probability of using words in the lexicon referentially are set to the maximum a posteriori values (the joint empirical Bayes estimate) to reduce the number of free parameters to one (the same number as the comparison models). After training, the model is scored both on the accuracy of its lexicon and on the accuracy of the inferences it makes about speakers' referential intentions. These scores are measured relative to a gold-standard lexicon and intention set generated by a human coder. The gold-standard lexicon included every noun (including plurals and baby talk, excluding pronouns) used to refer to an object at least once in the data. The gold-standard intents were based on Fernald & Morikawa's (1993) best guess as to the speakers' referents. The measures of accuracy used were *precision* (proportion of mappings made that were correct), *recall* (proportion of the total correct mappings that were found), and *F score* (the harmonic mean of precision and recall, commonly used as a standard measure of a model's degree of accuracy).

The model is compared against five other cross-situational word learning models to gauge its relative success, the first three of which are calculations of co-occurrence frequency, conditional probability, and point-wise mutual information. We also compared the Bayesian inference model to IBM Machine Translation Model I (Brown, Pietra, Pietra, & Mercer, 1994), computing association probabilities both for objects given words and words given objects. After a word-by-object matrix of association values was

attained for each model, a number of lexicons were created by considering a number of threshold values and only including word-object pairs with an association value higher than the threshold. The lexicon resulting from the threshold value that yielded the highest posterior score was kept for each model. The comparison models' intentional inferences for each situation were taken to be the objects for which a corresponding word in each model's best lexicon was uttered.

Each of the model's potential lexicons is scored based on its posterior probability, $p(L|C) \propto p(L) \times p(C|L)$, found by calculating the product of the prior and likelihood probabilities. The prior probability is calculated according to a parsimony assumption, awarding each lexicon L_i a score inversely proportional to its size:

$$p(L_i) \propto e^{-\alpha|L_i|}$$

The likelihood function, which calculates the probability $p(C|L_i)$ of observing the corpus of situations given a lexicon, is based on a number of interdependencies and assumptions. For the objects O_s , intentions I_s, and words W_s in each situation S, we assume that I_s is a subset of O_s, and that every subset is equally likely to be intentionally referred to, i.e., $p(I_s|O_s) \propto 1$. We further assume that given I_s, a speaker's utterance W_s depends upon both I_s and the lexicon L. We also assume that speakers have a certain probability γ of using a word referentially in any given context. We finally consider two distinct probabilities: firstly, the probability $p_R(w|o,L_i)$ of choosing a word $w \in W_s$ uniformly at random from the set of valid labels to refer to a given object $o \in O_s$ with lexicon L_i, and secondly, the probability $p_{NR}(w|L_i)$ of choosing a word to be used non-referentially. A parameter κ dictates how likely words in the lexicon are to be used non-referentially relative to words outside the lexicon (i.e., because we choose $\kappa < 1$, words in the lexicon are less likely to be used non-referentially). As our final likelihood probability we get:

$$p(C|L_i) = \prod_{S \in C} \sum_{I_s \subseteq O_s} \prod_{w \in W_s} \left[\gamma \cdot \sum_{o \in I_s} \frac{1}{|I_s|} p_R(w|o, L_i) + (1-\gamma) \cdot p_{NR}(w|L_i) \right]$$

Hypotheses for potential lexicons are generated stochastically: new lexicons are always chosen over old lexicons if they yield a greater posterior score, but are chosen with a probability equal to the ratio of the lexicon's scores otherwise. New lexicons are generated by adding a word-object pairing, deleting a pairing, or swapping two pairings according to a data-driven Markov-Chain Monte Carlo strategy. Because of the irregularity of the posterior score distribution, incremental moves in the right direction may actually temporarily yield severely worse posterior scores. The lexicon space is therefore searched stochastically via a Monte Carlo strategy known as simulated tempering whereby a number of searches with differing degrees of greediness are run in parallel. The model's search and scoring process typically converges to its final posterior value within 10k-50k moves.

Simulation 1

In the first simulation we extend the monolingual Bayesian word-learning simulation of Frank et al. (2009), using a new data set from similarly annotated transcriptions (Fernald & Morikawa, 1993) of English-speaking mothers interacting with their infants. This data set is compiled to provide a corpus comparable in size and complexity to the corpus used by Frank et al. (2009), see Table 1 for comparison.

Table 1

Size and Complexity Comparisons Between Past and Current Datasets

	Frank et al. (2009)	**Current Monolingual Study**	**Current Bilingual Study**
Size			
Object types	22	22	22
Word types	419	321	486*
Object tokens	1261	1671	1671
Word tokens	2507	2106	2019
Total situations	619	571	571
Complexity			
Average words per situation	4.0501	3.6883	3.5359
Average objects per situation	2.0372	2.9264*	2.9264*
Average words per object per situation	2.5987	1.5252*	1.4492*

Note. * indicates a significant difference from previous studies. In this case the presence of approximately one more object per situation on average produces a slightly more complex set of situations, predicting decreased performance.

Simulation 2

In the second simulation, the same model is trained on bilingual input. Approximately 50% of the utterances from the monolingual corpus were translated into Spanish, with transparency (i.e., the translation's Spanish nativeness) being chosen over fidelity (i.e., the extent to which the translation accurately renders the meaning of the English) whenever possible. Besides language differences, the bilingual corpus is similar in size and complexity to the corpus in Simulation 1, as can be seen in Table 1. The only significant difference between the corpora is the significantly larger (51%) number of word types in Experiment 2, a result that is expected from the use of two languages and by extension the common use of two words to designate the same concept. After training, the same accuracy ratings used in Simulation 1 are re-applied in Simulation 2.

Results

Simulation 1

The Bayesian intentional model was run 20 times, and its precision, recall, and *F* score were recorded after each run. After averaging these scores across all the runs, the results indicated that the model outperformed the comparison models in building a lexicon from the child-directed speech situations. As can be seen in Table 2, the intentional model had the highest precision value, .40, and the highest *F* score value, .35. Unlike the results in Frank et al. (2009), the intentional model did not have the highest recall score in our simulations; rather, the conditional probability model had the highest recall score at .59 as compared to the intentional model at .31.

Table 2

Precision, Recall, and F *Score of the best lexicon found by each model*

Model	Precision	Recall	*F* score
Association frequency	.04	.31	.07
Conditional probability (object|word)	.04	**.59**	.07
Conditional probability (word|object)	.29	.13	.17
Mutual information	.22	.34	.27
Translation model (object|word)	.15	.31	.20
Translation model (word|object)	.24	.38	.29
Intentional model	**.40**	.31	**.35**

Note. The highest values obtained are highlighted in boldface.

In contrast to findings in Frank et. al (2009), the model had no advantage in inferring speakers' intentions (Table 3), and did not have the highest value for any of the scores. The best precision was obtained by the mutual information model with a value of .56 (as compared to the intentional model's .26) and the best *F* score was obtained by the translation model with a value of .42 (as compared to the intentional model's .31). As in the previous study, the association frequency model obtained the highest recall value for the intentional inferences by a wide margin with a value of .75 (as compared to the intentional model's .39).

Table 3

Precision, Recall, and F *Score of the inferences made by each model about the speaker's referential intentions, using the lexicons scored in Table 2*

Model	Precision	Recall	*F* score
Association frequency	.14	**.75**	.24
Conditional probability (object|word)	.16	.61	.25
Conditional probability (word|object)	.50	.21	.30
Mutual information	**.56**	.18	.27
Translation model (object|word)	.55	.34	**.42**
Translation model (word|object)	.33	.51	.40
Intentional model	.26	.39	.31

Note. The highest values obtained are highlighted in boldface.

Table 4 displays the best lexicon found by the intentional model, which was considerably smaller than the best lexicons found by all but one other model (the translation model calculating the conditional probability of a word given an object was the only model to posit a smaller lexicon with size 14). Of the 26 lexical pairings posited by the intentional model, 13 were judged to be correct according to the gold standard. The remaining comparison models posited lexicons with sizes ranging from 50 to 500.

Table 4

Best Lexicon Found by Bayesian Intentional Model in Monolingual Simulations

Word	Object	Word	Object
hair	brush	you	face
flashlight	**flashlight**	**waffles**	**waffles**
dough	**dough**	under	pepperoni
doors	car	the	face
doggy	**dog**	the	dog
cheese	**cheese**	the	hotdog
brush	box	ruff	pig
brush	**brush**	**rosy**	**doll**
blocks	**blocks**	red	truck
bear	**bear**	**rabbit**	**rabbit**
bang	brush	leg	pepperoni
baby	**baby**	joey	book
alphabet	**alphabet**	**hotdog**	**hotdog**

Note. Word-object pairs judged to be correct according to the gold standard are highlighted in boldface.

Simulation 2

In Simulation 2, the Bayesian intentional model, while still performing highly competitively in determining a lexicon given bilingual speech situations (Table 5) only outperformed the comparison models in terms of precision with a value of .35. The association frequency model obtained the highest recall score with a value of .50 (as compared to the intentional model's .19) while the mutual information obtained the best *F* score with a value of .25 (fractionally beating out the intentional model's score). As expected, the model's best English and Spanish sub-lexicons, scored relative to the appropriate subset of the bilingual gold-standard, performed more poorly than the aggregate bilingual lexicon and more poorly than the monolingual lexicon.

Table 5

Precision, Recall, and F *Score of the best lexicon found by each model in a bilingual scenario*

Model	Precision	Recall	*F* score
Association frequency	.04	**.50**	.07
Conditional probability (object|word)	.04	.26	.07
Conditional probability (word|object)	.12	.17	.14
Mutual information	.22	.30	**.25**
Translation model (object|word)	.12	.26	.17
Translation model (word|object)	.15	.39	.22
Intentional model	**.35**	.19	.25
Intentional model (English only)	.21	.17	.19
Intentional model (Spanish only)	.25	.21	.23

Note. The highest values obtained are highlighted in boldface (differences between values may not be apparent because of rounding).

The model also performed rather poorly, both absolutely and relatively, in inferring the referential intentions of speakers (Table 6). For none of the three scoring metrics did the intentional model obtain the highest value. The best precision was, as in Study 1, obtained by the mutual information matrix with a value of .58 (as compared to the intentional model's .23) and the best *F* score was again obtained by one of the translation models with a value of .40 (as compared to the intentional model's .24). The association frequency model, as in Study 1 and in Frank et al. (2009), had the best recall score with a value of .82 (as compared to the intentional model's .25).

Table 6

Precision, Recall, and F *Score of the inferences made by each model about the speaker's referential intentions, using the lexicons scored in Table 5*

Model	Precision	Recall	*F* score
Association frequency	.15	**.82**	.25
Conditional probability (object|word)	.13	.27	.18
Conditional probability (word|object)	.43	.31	.36
Mutual information	**.58**	.17	.26
Translation model (object|word)	.48	.27	.34
Translation model (word|object)	.34	.49	**.40**
Intentional model	.23	.25	.24

Note. The highest values obtained are highlighted in boldface.

Table 7 displays the best bilingual lexicon found by the intentional model. The intentional model posited the smallest lexicon, again of size 26, with a total of 10 correct pairings. The comparison models posited lexicons with significantly more word-object pairings, ranging from 60 up to 600.

Table 7

Best Lexicon Found by Bayesian Intentional Model in Bilingual Simulations

Word	Object	Word	Object
you	face	**hotdog**	**hotdog**
you	box	guau	dog
wha	flashlight	grande	bear
the	face	**gofres**	**waffles**
the	dog	**doggy**	**dog**
the	hotdog	cheese	pepperoni
ruff	pig	cepilla	box
rosy	**doll**	**car**	**car**
queso	**cheese**	**brush**	**brush**
over	rabbit	**bloques**	**blocks**
oh	face	**bebe**	**baby**
maza	**dough**	bang	box
joey	book	a	face

Note. Word-object pairs judged to be correct according to the gold standard are highlighted in boldface.

Discussion

The use of a new monolingual English dataset yielded surprisingly different results from those obtained in Frank et al. (2009). While a number of the corpora's size metrics are similar, the smaller word token count combined with the larger object token count introduced additional ambiguity about the intended referents of each word in a situation, as compared to previous simulations. This increased complexity may have allowed the formation of spurious word-object pairs, as positive examples were less certain under the increased noise in the input. Consistent with previous results, the best lexicon found by the Bayesian model was still significantly smaller than those generated by the comparison models, owing to the bias of the prior likelihood toward parsimony. However, Table 7 reveals that, unlike the best lexicon from Frank et. al (2009), the best lexicon found in the present study contained a large number of spurious lexical items (e.g., the high frequency word 'the' was paired with the object 'dog', and the high frequency object 'face' was paired with the word 'you') despite the model's distinction between referential and nonreferential words and its bias to expect the latter. As one might expect from the low precision of the lexicon, the Bayesian model's intentional inferences based on the lexicon decreased in performance as well. Because bilingual infants have a number of other tools when discerning word meanings in realistic word learning situations (e.g., phonological, prosodic, lexical co-occurrence knowledge), in future studies we will identify the impact of removing words with obvious referents (e.g., 'you') or obvious non-referents (e.g., 'the') from the corpus, as children would most likely already have ruled these words out of the lexicon by other means.

In the next simulation, we assessed how a similarly complex but bilingual scenario affected the model's performance in building lexicons and inferring intentions when all else is held constant.

Besides a few exceptional cases, the bilingual data set yielded poorer performance than the monolingual data set for all models, and in most measures the Bayesian model was out-performed by the competing models. The best lexicon derived by the Bayesian model in this simulation contained even more obviously spurious lexical items than that of the previous simulation. Interestingly, both lexicons also contained several many-to-one word-object pairs along with a number of one-to-many (and in some cases many-to-many) word-object pairs. While any one-to-many word-object pairing is necessarily incorrect, as no two objects in the dataset have the same name, many-to-one pairs are of interest in assessing the degree to which the mutual exclusivity constraint is adhered to by the model.

As opposed to the results of Frank et. al (2009), in no case did the lexicons in either of these studies make a many-word-to-single-object pairing that was correct. While this may have been expected in the monolingual case where word learners have been found to rely heavily on the mutual exclusivity constraint (though the appearance of incorrect many-to-one pairings challenges this hypothesis), the lack of correct many-to-one pairings in the bilingual case, where the data was intended to catalyze this very type of violation, is highly problematic. The presence of these pairings in Frank et al. (2009) suggests, however, that the absence of these pairings in the current studies may be linked more to the roots of the drastic performance differences between Simulation 1 and Frank et. al (2009) than differences between learning from monolingual and bilingual inputs.

An analysis of the precision, recall, and *F* scores obtained when we score the bilingual lexicon within-language reveals some psychologically realistic results. To score the bilingual's solely English performance, for example, we removed the correct Spanish pairings made by the model from the lexicon and compared the remaining pairs (i.e., correct English pairings and any incorrect pairings) to the English subset of the gold-standard lexicon. In accordance with the literature, the bilingual's strictly English lexicon is smaller and less accurate than that of the monolingual English model (Bialystok et al., 2010). As one would expect though, the number of correct word-object pairings learned by the monolingual model and aggregate bilingual model are roughly the same, indicating similar vocabulary development trajectories.

For the most part the comparison models seem to have outperformed the Bayesian model in these more complex situations because they are unconstrained by any assumption of parsimony, and as a result they may make better use of a weak signal-to-noise ratio in the input. This advantage may decrease as the size of the corpus increases because the Bayesian model should improve in precision by excluding spurious non-referential words while the comparison models cannot do this.

Overall these simulations reveal an extreme sensitivity on the part of all models, and the intentional model in particular, to small changes in the training data. In short, our simulations in this study indicate that while the monolingual model performed comparably to the Frank et al. (2009) study, the model's absolute and relative performance was attenuated with increased complexity. While some work has here been done to define a principled way to track differences between transcriptions and datasets through comparisons of a number of size and complexity metrics, more must be done if model performance is to be reliably compared across corpora. While it appears that the Bayesian model displayed attenuated performance in the bilingual case, because of the

disparities between previous and current monolingual scores the degree and causes of this attenuation are questions for further research.

References

Aslin, R. N., & Newport, E. L. (2012). Statistical learning : From acquiring specific items to forming general rules. *Distribution*.

Au, T. K., & Glusman, M. (1990). The principle of mutual exclusivity in word learning : To honor or not to honor? *Child Development*, *61*(5), 1474-1490.

Bialystok, E., Barac, R., Blaye, A., & Poulin-Dubois, D. (2010). Word mapping and executive functioning in young monolingual and bilingual children. *Journal of cognition and development : official journal of the Cognitive Development Society*, *11*(4), 485-508.

Bonawitz, E. B., & Griffiths, T. L. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models.

Brysbaert, M., Verreyt, N., & Duyck, W. (2010). Models as hypothesis generators and models as roadmaps. *Bilingualism: Language and Cognition*, *13*(03), 383-384.

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: Infants' language experience influences the development of a word-learning heuristic. *Developmental science*, *12*(5), 815-23

Clark, A. (1989). Microcognition: Philosophy, cognitive science, and parallel distributed processing. Cambridge, MA: MIT Press.

Clark, E. V. (1993). *The lexicon in acquisition*. New York: Cambridge University Press.

Davidson, D., & Tell, D. (2005). Monolingual and bilingual children's use of mutual exclusivity in the naming of whole objects. *Journal of experimental child psychology*, *92*(1), 25-45.

Davis, S. J., Newport, E. L., & Aslin, R. N. (2009). Probability-matching in 10-month-old infants. *Methods*.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2008). Supplementary material for "Using speakers' referential intentions to model early cross-situational word learning." *Machine Translation*, 1-16.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 1-8.

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

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357-364. Elsevier Ltd.

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive psychology*, *63*(2), 93-106.

Jackson, F., and Pettit, P. (1988). Functionalism and broad content. *Mind, 97(387)*: 381-400.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Science*, *83*, 4-5.

Li, P. (2009). Lexical organization and competition in first and second languages: Computational and neural mechanisms. *Cognitive science*, *33*(4), 629-64.
Li, P., Burgess, C., & Lund, K. (2000). The acquisition of word meaning through global lexical co-occurrences. *Young Children*.
Li, P., & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. In R. Heredia & J. Altarriba, Bilingual sentence processing, 17, 59–85. North-Holland.
Li, P., & Zhao, X. (2012a). Connectionist approaches to bilingual lexical representation. In J. Altarriba & R. Heredia (eds.), Understanding bilingual memory: Theory and application. Springer Science Publishers.
Li, P., & Zhao, X. (2012b). Connectionism. In M. Aronoff (ed.), *Oxford bibliographies online.* New York, NY: Oxford University Press. (www.oxfordbibliographies.com)
Li, P., Zhao, X., & Mac Whinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive science*, *31*(4), 581-612.
Liittschwager, J. C., & Markman, E. M. (1994). Sixteen- and 24-month-olds' use of mutual exclusivity as a default assumption in second-label learning. *Developmental Psychology*, *30*(6), 955-968.
Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, *92*, 199-227.
Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *157*, 121-157.
Marr, D. C. (1982). Vision. San Francisco: Freeman.
Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. Psychological Review.
Paradis, J. (2001). Beyond 'one system or two?': Degrees of separation between the languages of French-English bilingual children. In S. Dopke (Ed.), *Cross-linguistic structures in simultaneous bilingualism* (pp. 175–200). Amsterdam: Benjamins.
Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302-321. Elsevier B.V.
Quine, W. V. O. (1960). Word and object. Cambridge, MA: MIT Press.
Richardson, F. M., & Thomas, M. S. (2008). Critical periods and catastrophic interference effects in the development of self-organizing feature maps. Developmental science, 11 (3), 371-89.
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.
Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds : a computational model. *Cognitive Science*, *26*, 113-146.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.
Seidenberg, M. S., & Mcclelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523-568.
Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39-91.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-68.
Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1536), 3617-32.
Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729-42.
Wittgenstein, L. (1922). Tractatus logico-philosophicus. New York, NY: Routledge.
Xu, F., & Tenenbaum, J. B. (2007a). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245-72.
Xu, F., & Tenenbaum, J. B. (2007b). Sensitivity to sampling in Bayesian word learning. *Developmental science*, *10*(3), 288-97.
Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, *4*(1), 32-62.
Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, *18*(5), 414-20.
Yurovsky, D., Fricker, D., Yu, C., & Smith, L. B. (2010). The active role of partial knowledge in cross-situational word learning. *Cognitive Science*.
Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. *Cognitive Science*.
Zhao, X., & Li, P. (2007). Bilingual lexical representation in a self-organizing neural network. Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum.
Zinszer, B. D., & Li, P. (2011). A SOM model of first language lexical attrition, 2787-2792.