Robust Lexical Acquisition Despite Extremely Noisy Input

Jeffrey Mark Siskind, University of Toronto

1 Introduction

Noise is a central problem facing a language learner. Any theory of language acquisition must explain how children robustly make correct categorical decisions about their native language even though an unmarked portion of the primary linguistic data is ungrammatical. Lexical acquisition is particularly plagued by noise. While perhaps only a small percentage of the utterances heard by children are ungrammatical, the correlation between word and world may be much more tenuous. For instance, Gleitman (p.c.) reports that opening events occur less than 70% of the time that children hear the word *open* and that the vast majority of the time that openings occur, the word *open* isn't even uttered. This raises the obvious question: How can a child determine that *open* means OPEN when, on the surface, much of the evidence suggests otherwise. The problem of noisy input has motivated some authors (e.g. Gleitman 1990, Fisher et al. 1994) to suggest that lexical acquisition based solely on word-to-world correspondences is impossible and to conjecture alternative strategies that use syntactic information to guide acquisition. Such strategies have become known as *syntactic bootstrapping*.

A child might learn a word by hearing it in several different contexts and deciding that it means something that is invariant across those different contexts. For instance, a child hearing *John lifted the ball*, while seeing John lift a ball, and *Mary lifted a box*, while seeing Mary lift a box, might determine that *lifted* refers to the lifting event, and not John, Mary, the ball, or the box, since the latter do not remain invariant across the two events. This general strategy has been proposed by numerous authors. For instance, Gleitman and Fisher et al. call this procedure *cross-situational learning* while Pinker (1989) calls it *event category labeling*. Siskind (1994) and Siskind (to appear) present a precise formulation of a procedure based on this strategy.

The cross-situational strategy suffers from a fundamental flaw, however. What happens when a child hears an utterance that contains the word *lift* when no lifting occurs? In this case, there will be no potential referent that is invariant across all uses of the word *lift*. I refer to such utterances as *noise*. In the more general case, where utterances are paired with sets of hypothesized meanings, an utterance is considered to be noisy if all of the hypothesized meanings are incorrect. The main purpose of this paper is to present a strategy for learning word meanings even in cases where as many as 90% of the utterances heard by the learner are noisy.

In this paper, I present a precise implemented algorithm capable of acquiring a lexicon of word-to-meaning mappings from input similar to that available to children. An important characteristic of this algorithm is that it can acquire such a lexicon with greater than 95% accuracy despite the fact that over 90% of the input

is noisy. It does so *without* using any syntactic information to guide the acquisition process, thus suggesting that inferences based on the syntactic structure of utterances might not be strictly necessary for successfully acquiring word meanings. The algorithm achieves this performance by means of a cascade of two processes, one making use of statistical correlations and the other applying more categorical constraints. The statistical process consists of a set of linear equations that relate two sets of variables, one characterizing the semantic contribution of each word in the lexicon and the other measuring the expected semantic token occurrence rate conditional on word occurrence. These equations constitute a model of the underlying noise generation process under a number of weak assumptions. By solving these equations, one can get an estimate of the semantic contribution of each word (i.e. the unknown lexicon) from the observed semantic token occurrence rates.

The statistical process itself is not robust. The accuracy of the lexicon it produces degrades significantly as the noise rate increases beyond 70%. Nonetheless, the results of the statistical process can be used to predict which subsequent utterances are likely to be noise. Thus it can be used as an input filter to a second, more categorical process. For this, the statistical process need only be sufficiently accurate to reduce the noise rate to levels that can be tolerated by the categorical process without discarding too much of the data. In the remainder of this paper, I describe the algorithm in greater detail and present the results of experiments that demonstrate that it is capable of reliably learning small lexica from noisy synthetic corpora that are of different sizes and that exhibit different noise rates.

I should state at the outset that I do *not* claim that children actually use any of the techniques that I present in this paper. This paper merely investigates the capabilities and limitations of one possible approach that children *might* employ as part of their lexical acquisition strategy. This approach differs in many ways from those normally explored within the child language acquisition research community. Further experimental evidence might help determine what role, if any, the techniques described in this paper play in actual child language acquisition.

2 The Formal Problem

When learning their native language, children must learn a lexicon that maps words to representations of their meanings. For instance, children learning English must learn that *open* refers to opening events while *door* refers to doors. The task of learning such word-to-meaning mappings has become known as the *mapping problem*. The key difficulty in this task is determining, from a multi-word utterance, which words map to which meanings. For example, when hearing the utterance *The door opened*, how can the child determine that *open* refers to the opening event, while *door* refers to the door, and not vice versa?

Children must, of course, solve numerous other problems during lexical acquisition besides the mapping problem. For instance, not only must they determine

what words mean, they must also determine which strings of sounds constitute words in the first place. Additionally, they must learn the possible morphological variation to words and what semantic features these variations encode. Furthermore, they must learn a mapping from words to parts of speech and, for words that take arguments, the allowed syntactic forms for realizing those arguments. Other authors (e.g. Grimshaw 1979, Pinker 1989, Marcus et al. 1992, Brent et al. 1994) have addressed many of these learning problems. This paper focuses solely on the problem of learning word-to-meaning mappings.

Let us adopt a simple model of the mapping problem. Suppose that children hear a sequence of utterances, each being a sequence of words. Furthermore, let us suppose that when hearing an utterance, children can correctly determine the utterance meaning from context. This is, of course, a rather strong assumption. I will relax this assumption momentarily. Given this assumption, however, solving the mapping problem involves breaking the meanings of whole utterances into parts and assigning those parts as the meanings of individual words.

As stated above, the mapping problem is under-constrained. One can adopt any possible mapping between the words and meaning fragments of each utterance independently from the mapping adopted for other utterances. Doing so could map a given word to different meanings in different utterances. For example, upon hearing *The door opened*, while seeing a door open, the learner could map *door* to OPEN and *open* to DOOR. Later, upon hearing *The door closed*, while seeing a door close, the learner could map *door* to DOOR and *close* to CLOSE, thus obtaining two different mappings for the word *door*. To preclude this possibility, I assume that the learner adopts a *monosemy* constraint, namely the default assumption that each word must have at most one meaning. Again, this assumption is, of course, too strong. It serves only as a default assumption and is relaxed later in this paper. It is interesting to point out that, when one adopts a monosemy constraint, almost all instances of the mapping problem have a unique solution, if they have a consistent solution at all, so long as there is a sufficiently large ratio between the number of utterances in the corpus and the vocabulary size.

Some authors have proposed a converse constraint prohibiting synonyms instead of homonyms. Such a constraint requires each meaning to map to one word instead of requiring each word to map to one meaning. The learning algorithms that I present in this paper do not prohibit synonyms.

The model described so far makes three overly-restrictive assumptions: that the learner can always determine the correct utterance meaning from context, that each word in the lexicon has a single meaning, and that the correct meaning of each utterance can always be derived from the meanings of its constituent words. I relax each of these assumptions by making two extensions to the model. First, instead of requiring the learner to hypothesize a single correct meaning for each utterance from context, I allow the learner to hypothesize a *set* of possible meanings for an utterance. For example, when hearing an utterance like *Mommy lifted the ball*, while seeing Mommy lift a ball, the learner might guess that this utterance meant

that Mommy lifted the ball, that she grasped the ball, or even that she wanted the ball. Only one element of this set (the first in this case) is correct. The remainder are spurious. If the learner hypothesizes a set of meanings for an utterance, instead of a single meaning, the learning problem becomes confounded. The learner must not only determine how to decompose an utterance meaning and assign its constituents to words in the utterance, she must also determine which hypothesized utterance meaning to decompose in the first place. I refer to this additional level of ambiguity in the learning process as *referential uncertainty*.

Even allowing for referential uncertainty is not sufficient. What happens when *none* of the referentially uncertain hypothesized utterance meanings are correct? This might happen when an utterance does not refer to the immediate context. Or it might happen when an utterance is ungrammatical, since there might be no way to form the meaning of such an ungrammatical utterance from the meanings of its constituent words in a way that is consistent with the semantic interpretation rules of the language. Finally, this might happen when an utterance contains a homonymous word. The learner might have already learned one sense for some word yet hear an utterance that can map to one of the hypothesized meanings only if that word is allowed to take on a new sense. I refer to all such situations collectively as *noise*. A certain fraction of the input corpus will be noisy and should be ignored. The learner, however, does not know *a priori* which utterances are noisy, for there is no simple way that such utterances are marked in the input. The main contribution of this paper is a method by which a learner can gather information during early stages of lexical acquisition to help filter out noise during later stages of that process.

3 Semantic Representations

In order to present a precise algorithmic solution to the mapping problem, it is necessary to specify the form taken by semantic representations for the meanings of words and utterances. While numerous authors (e.g. Leech 1969, Miller 1972, Jackendoff 1983, and Pinker 1989) have pursued theoretical investigations into the possible nature of semantic representation, there is little concrete evidence detailing the precise form of human mental representations. Accordingly, in this paper, I adopt a minimal set of assumptions that delineate the representations used for words and utterances. First, I assume that there is an inventory of *conceptual symbols* used to form meaning representations. That inventory might include symbols like GO, CAUSE, UP, **John**, and **ball**. Second, I assume that these conceptual symbols are used to form *expressions* to denote the meanings of words and utterances. So, for instance, the meanings of the words *John*, *lifted*, and *ball* might be represented as the expressions **John**, CAUSE(x, GO(y, UP)), and **ball** respectively, while the meaning of the whole utterance *John lifted the ball* might be represented as the expression CAUSE(**John**, GO(**ball**, UP)).

The algorithms that I describe in this paper do not presuppose a particular

inventory of conceptual symbols or a particular way of combining such symbols to form meaning expressions. They should work for any way of representing word and utterance meanings as expressions, including those of Leech, Miller, Jackendoff, and Pinker. Furthermore, meaning representations are treated as expressions over uninterpreted symbols. The acquisition of word-to-meaning mappings is viewed simply as the process of pulling utterance meaning representations apart and assigning fragments of those representations as the meanings of the words in the utterance.

Note that there is no requirement that each word map to an expression containing a single conceptual symbol. For example, the word *lift* might map to the expression CAUSE(x, GO(y, UP)) which contains three conceptual symbols. It is even possible that the meaning representation for some word might contain more than one occurrence of the same conceptual symbol. For example, the expression CAUSE(GO(x, UP), GO(y, UP)) might be a more detailed representation of the word *lift*.

4 A Statistical Model of Noise

The lexical acquisition algorithm that is described here operates in two stages. The first stage learns the set of conceptual symbols used to construct the meaning representation for a word. The second stage then learns how to assemble these symbols into an aggregate meaning expression. For example, the first stage would learn that the meaning of the word *lift* contains the conceptual symbols CAUSE, GO, and UP. The second stage would then learn that the proper way to combine these symbols to represent the meaning of *lift* is CAUSE(x, GO(y, UP)), and not for instance, UP(CAUSE, GO(x, CAUSE, x)) or CAUSE(x, GO(x, UP)).

Tishby and Gorin (1994) present one method for performing the first stage. They construct three matrices, A, B, and C, to represent the lexicon, the set of training utterances, and the hypothesized meanings of the training utterances respectively. A_{kj} denotes the number of times the conceptual symbol j appears in the meaning representation of word k. B_{ik} denotes the number of times the word k appears in utterance i. C_{ij} denotes the number of times the conceptual symbol j appears in the hypothesized meaning representation for utterance i. They further assume that each utterance is paired with a single hypothesized meaning, that each word maps to a single meaning, and that the meaning of each utterance contains precisely the union of those conceptual symbols that appear in the meanings of the words that make up that utterance. In other words, if an utterance contained the words *Mommy*, *lifted*, *the*, and *ball*, and these words contribute the following conceptual symbol sets respectively: {**mother**}, {CAUSE, GO, UP}, {}, and {**ball**}, then the meaning of the whole utterance must contain precisely the union of these sets, namely {**mother**, CAUSE, GO, UP, **ball**}. Thus they assume that there is no referential uncertainty, homonymy, or noise, and that the semantic interpretation rule cannot add, delete, or copy information when composing the

meaning of an utterance from the meanings of its parts. Given these assumptions, Tishby and Gorin observe that $C = BA$, and that since B and C are observable from the corpus, the hidden lexicon A can be recovered by computing $B^{-1}C$.

The first stage of the lexical acquisition algorithm that is described here is divided into two sub-stages, the first a statistical process and the second a more categorical process. The statistical process is similar, in many ways, to the algorithm proposed by Tishby and Gorin except that it allows for referential uncertainty and noise. The subsequent categorical process handles homonymy and relaxes the restriction on the semantic interpretation rule to allow copying.

In the statistical process, $\hat{R}(w, f)$ denotes the average number of occurrences of the conceptual symbol f in the set of hypothesized utterance meanings associated with utterances that contain at least one occurrence of the word w. Further, η denotes the noise rate, the fraction of utterances paired only with incorrect utterance meanings. $\hat{K}$ denotes the average degree of referential uncertainty, the average number of hypothesized utterance meanings paired with each utterance. $\hat{J}$ denotes the mean utterance length (MLU), the average number of words in an utterance. W denotes the vocabulary, the set of all words that appear in the training corpus. I use $o(w)$ to denote the number of times the word w appears in the corpus. Finally, $Q(w, f)$ denotes the number of occurrences of the conceptual symbol f in the representation of the meaning of the word w.

The quantities $\hat{R}(w, f)$, $\hat{K}$, $\hat{J}$, W, and $o(w)$ are measurable from the corpus. $Q(w, f)$ constitutes both the mental lexicon used by the speaker when producing the utterances, as well as the mental lexicon to be constructed by the learner. For instance, if the meaning of *lift* is CAUSE(x, GO(y, UP)) then $Q(lift, \text{CAUSE}) = 1$, $Q(lift, \text{GO}) = 1$, and $Q(lift, \text{UP}) = 1$, while $Q(lift, x) = 0$ for $x \neq \text{CAUSE}$, $x \neq \text{GO}$, and $x \neq \text{UP}$. Note that $Q(w, f)$ can be greater than one if the representation of a word meaning contains more than one instance of some conceptual symbol. While the underlying true, but hidden, $Q(w, f)$ will be integral, the recovered estimation of $Q(w, f)$ might be nonintegral.

$Q(w, f)$ constitutes a representation of word meanings while $R(w, f)$ constitutes a representation of the hypothesized utterance meanings. Relating these two quantities requires some assumptions about the semantic interpretation process, namely, how word meanings combine to form utterance meanings. For the statistical first stage of the lexical acquisition process, I adopt the same assumptions as Tishby and Gorin. More specifically, I assume that the number of times a particular conceptual symbol appears in the meaning of an utterance must equal the sum of the numbers of times that symbol appears in the meanings of words in that utterance. This semantic interpretation rule is overly restrictive. It requires that all semantic information in an utterance derive from words in that utterance and not, say, the syntactic form of an utterance. It also rules out deletion or duplication of semantic material. The fact that real language might exhibit such phenomena, however, does not preclude the use of the algorithm presented here. Such phenomena might occur only in some, but not all, of the utterances in a

typical training corpus. These utterances can be treated as noise by the lexical acquisition process. Furthermore, only the statistical first stage of the process that I describe makes such stringent assumptions. Later stages of the process relax many of these restrictions.

We can now derive a prediction for $\hat{R}(w,f)$ given the remaining parameters. This constitutes a generative model for how the corpus was produced. Each utterance is either noise or is paired with a correct meaning. The former occurs with frequency η while the latter with frequency $1-\eta$. First consider the case of noisy utterances. In this case, the learner is presented with $\hat{K}$ meaning representations, on the average. Let us make two assumptions as to how these meaning representations are generated. First, let us assume that they correspond to linguistically realizable utterances. In other words, learners hypothesize as potential meanings for a given utterance only those expressions that could correspond to some utterance. Second, let us assume that the expected length of such hidden utterances is equal to the observed MLU of the training corpus and that the words in these utterances are selected independently with the same frequency as observed in the corpus. While these assumptions are clearly false, they are adequate approximations for our purposes.

Given these assumptions, each of the $\hat{K}$ incorrect meaning representations paired with a noisy utterance corresponds to a hidden utterance containing, on the average, $\hat{J}$ words. Each such word is likely to be the particular word w_i with the following frequency:

$$\frac{o(w_i)}{\sum_{w \in W} o(w)}$$

Since w contributes $Q(w,f)$ instances of the conceptual symbol f, the expected number of instances of f among all of the incorrect meaning representations associated with a noisy utterance is the following:

$$\hat{K}\hat{J}\left(\frac{\sum_{w \in W} o(w)Q(w,f)}{\sum_{w \in W} o(w)}\right)$$

Now let us consider the case where an utterance is paired with a correct meaning. In this case, there are, on the average, $\hat{K}$ meanings hypothesized for the utterance. One of these must be correct. Let us assume that the remaining $\hat{K}-1$ are generated by the same process that generates meaning representations for noisy utterances. Thus the expected number of instances of f among these $\hat{K}-1$ hypothesized meanings is the following:

$$(\hat{K}-1)\hat{J}\left(\frac{\sum_{w \in W} o(w)Q(w,f)}{\sum_{w \in W} o(w)}\right)$$

This leaves the issue of how the correct utterance meaning is produced. Let us assume that this meaning is produced by the same process as all of the others. Thus this utterance meaning corresponds to a hidden utterance containing $\hat{J}$ independently selected random words. Recall that we are interested in computing $\hat{R}(w,f)$, the average number of occurrences of the conceptual symbol f in meanings associated with an utterance that contains *at least one* occurrence of w. That word must contribute $Q(w,f)$ instances of f. The remaining $\hat{J}-1$ words will, on the average, contribute

$$(\hat{J}-1)\left(\frac{\sum_{w\in W} o(w)Q(w,f)}{\sum_{w\in W} o(w)}\right)$$

instances of f.

Thus, overall, the expected number of instances of f in the set of hypothesized meaning expressions associated with an utterance that contains w is given by the following formula:

$$\begin{aligned}\hat{R}(w,f) &= (1-\eta)Q(w,f) + \\ &\quad [\eta\hat{K}\hat{J} + (1-\eta)(\hat{K}-1)\hat{J} + (1-\eta)(\hat{J}-1)]\left(\frac{\sum_{w\in W} o(w)Q(w,f)}{\sum_{w\in W} o(w)}\right)\end{aligned}$$

This can be thought of as a generative model explaining how the corpus was created given the parameters η, $\hat{K}$, $\hat{J}$, W, $o(w)$, and $Q(w,f)$. These parameters reside collectively in the head of the speaker, who chose which utterances to say, and in the head of the hearer, who chose which meanings to hypothesize for those utterances. The goal of lexical acquisition is to recover the hidden $Q(w,f)$ given the remaining observable parameters of the corpus.[1]

Let $\hat{R}(f)$ denote the vector of values $\hat{R}(w,f)$ for all $w \in W$. Similarly, let $Q(f)$ denote the vector of values $Q(w,f)$ for all $w \in W$. Given this, it is possible to formulate the above relation between $\hat{R}(w,f)$ and $Q(w,f)$ as a set of linear equations $\hat{R}(f) = AQ(f)$ where:

$$A = \begin{bmatrix} \alpha_1+\beta & \alpha_2 & \cdots & \alpha_n \\ \alpha_1 & \alpha_2+\beta & \cdots & \alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n+\beta \end{bmatrix}$$

$$\alpha_i = [\eta\hat{K}\hat{J} + (1-\eta)(\hat{K}-1)\hat{J} + (1-\eta)(\hat{J}-1)]\frac{o(w_i)}{\sum_{w\in W} o(w)}$$

$$\beta = 1-\eta$$

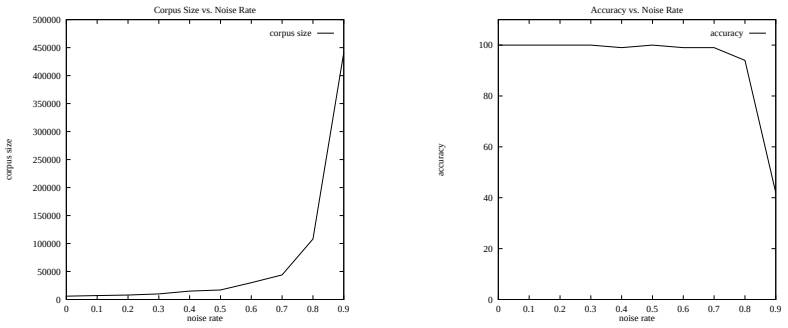
Thus the learner can estimate the hidden values for $Q(w, f)$ simply by computing $Q(f) = A^{-1}\hat{R}(f)$. Fortunately, there is a closed-form representation for A^{-1}:

$$
\begin{aligned}
A^{-1} &= \frac{1}{\beta(\beta + \sum_{i=1}^{n} \alpha_i)} \begin{bmatrix} \delta_1 & \gamma_2 & \cdots & \gamma_n \\ \gamma_1 & \delta_2 & \cdots & \gamma_n \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_1 & \gamma_2 & \cdots & \delta_n \end{bmatrix} \\
\gamma_j &= -\alpha_j \\
\delta_i &= \beta + \sum_{k=1, k \neq i}^{n} \alpha_k
\end{aligned}
$$

5 Experiments

It is impossible to test this technique on real corpora of adult speech to children since no such corpora exist that have been annotated with semantic information. Thus I have tested it on synthetic corpora randomly generated with a variety of distributional parameters controlling vocabulary size, mean utterance length, degree of referential uncertainty, size of conceptual vocabulary, complexity of conceptual expressions, noise rate, and so forth. In one series of experiments, a base-line set of parameter values was chosen and then the noise rate was varied from 0% to 90%, measuring the corpus size needed to acquire the meanings of all words in that corpus with 95% accuracy. For these experiments, the vocabulary size was set at 100 words, the MLU was approximately 5, the degree of referential uncertainty was 10 meanings per utterance, the conceptual vocabulary included 25 symbols, and conceptual expressions denoting the meanings of whole utterance could contain up to 30 symbols. Figure 1(a) illustrates the requisite corpus size, in number of utterances, to achieve 95% lexical acquisition accuracy as a function of the noise rate. Another series of experiments was performed, with the same base-line parameters, that fixed the corpus size at 100,000 utterances and measured the lexical acquisition accuracy as a function of the noise rate. Figure 1(b) illustrates the results of this second series of experiments.

These experiments demonstrate that the lexical acquisition algorithm works well for noise rates as high as 70%. The accuracy of the acquired lexicon, however, degrades rapidly for higher noise rates. Higher noise rates require larger training corpora to get robust estimates of $\hat{R}(w, f)$ and $Q(w, f)$. Recall that $\hat{R}(f) = AQ(f)$ and $Q(f) = A^{-1}\hat{R}(f)$. A is contractive. In other words, large differences in $Q(f)$ correspond to small differences in $\hat{R}(f)$. On the other hand, A^{-1} is expansive. Small differences in the measured values for $\hat{R}(f)$ result in large differences in the estimates recovered for $Q(f)$. The dependency on the noise rate is apparent in the equations for A and A^{-1}. A becomes singular as $\beta \to 0$. Since $\beta = 1 - \eta$, A becomes singular as $\eta \to 1$. In other words, the algorithm breaks down when the input consists solely of noise. This is not surprising. It is nonetheless quite



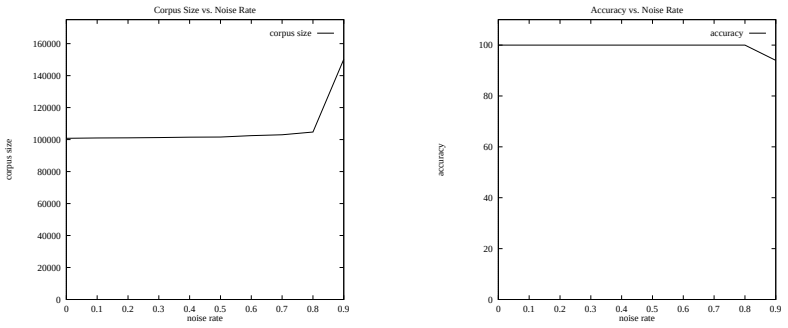
(a) (b)

Figure 1: (a) The requisite corpus size, in utterances, needed to achieve 95% lexical acquisition accuracy, using only stage one, as a function of the noise rate. (b) Lexical acquisition accuracy, using only stage one, as a function of the noise rate for a fixed-size corpus of 100,000 utterances.

encouraging that this technique works as well as it does for noise rates as high as 70%, without requiring excessively large corpora.

It is possible, however, to get even better performance at even higher noise rates. Siskind (1994) and Siskind (to appear) present a more categorical process for acquiring word-to-meaning mappings. This process also handles referential uncertainty and noise. In addition, it handles homonymy, makes fewer assumptions about the semantic interpretation process, and learns how to combine conceptual symbols to form conceptual expressions. Thus it learns not only that *lift* contains CAUSE, GO, and UP in its meaning but also that these symbols are arranged as the expression CAUSE(x, GO(y, UP)). This process suffers from a shortcoming however. While it can robustly learn a lexicon with low levels of noise—under 20%—it quickly breaks down with higher levels of noise.

This suggests the following possibility. The statistical algorithm described in this paper can be used as the first stage of a two-stage process. The categorical algorithm can be used as the second stage. In the first stage, the learner listens to a portion of the corpus and measures $\hat{R}(w,f)$, $\hat{K}$, $\hat{J}$, W, and $o(w)$. After listening to a sufficiently large sample to robustly measure these quantities, the learner computes $Q(w,f)$. The learner's estimate of $Q(w,f)$ will be inaccurate. Nonetheless, it can be used to predict whether or not future utterances are noisy. The learner can then begin the second stage, processing the remainder of the corpus using information gathered in the first stage as a noise filter. While $Q(w,f)$ will not be sufficiently accurate to correctly distinguish noisy utterances from good ones 100% of the time, it is not necessary to do so. The filter need only reduce the noise rate to levels that the second categorical process can deal with. It can erroneously pass through some noisy utterances—and filter out some good ones—so long as it



(a) (b)

Figure 2: (a) The requisite corpus size, in utterances, needed to achieve 95% lexical acquisition accuracy, using both stages, as a function of the noise rate. (b) Lexical acquisition accuracy, using both stages, as a function of the noise rate for a fixed-size corpus of 150,000 utterances.

doesn't over-zealously filter out too many of the good utterances.

A third series of experiments was performed, with the same base-line parameters as before, using both the statistical first stage and the categorical second stage. In this series of experiments, the cut-over from the first stage to the second was fixed at the 100,000th utterance. Figure 2(a) illustrates the corpus size needed to achieve 95% lexical acquisition accuracy as a function of the noise rate, while figure 2(b) illustrates the lexical acquisition accuracy as a function of the noise rate for a fixed-size corpus of 150,000 utterances.

6 Conclusion

I have demonstrated a novel two-stage method for acquiring word-to-meaning mappings that combines a statistical first stage with a categorical second stage. While neither of these processes alone is capable of robustly acquiring a lexicon in the presence of high noise rates, together they can achieve greater than 95% acquisition accuracy despite as much as 90% noise. Neither the statistical first stage nor the categorical second stage make any use of word order. Thus they suggest that, at least in theory, children might not need to use syntactic information to guide the acquisition of word-to-meaning mappings, at least for the bulk of the lexicon. While I do not claim that children employ the techniques described in this paper, they could—in principle—do so. Child language acquisition is a complex process. Children might use a multitude of techniques, each in a different situation or at a different stage. Further investigation is necessary to determine what role, if any, the techniques described in this paper play in that process.

Notes

∗ Peter Dayan provided considerable assistance developing the statistical model described in this paper, and in particular, helping find the closed-form inverse for *A*. Any errors in this paper, of course, are the sole responsibility of the author. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

1. Technically, η, the noise rate, is also hidden. Future work will address the question of how to estimate this parameter as well.

References

Brent, M. R., Gafos, A., and Cartwright, T. A. (1994). Phonotactics and the lexicon: Beyond bootstrapping. In Clark, E., editor, *Proceedings of the 1994 Stanford Child Language Research Forum*. Cambridge University Press, New York, NY.

Fisher, C., Hall, G., Rakowitz, S., and Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92(1):333–375.

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

Grimshaw, J. (1979). Complement selection and the lexicon. *Linguistic Inquiry*, 10:279–326.

Jackendoff, R. (1983). *Semantics and Cognition*. The MIT Press, Cambridge, MA.

Leech, G. N. (1969). *Towards a Semantic Description of English*. Indiana University Press.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., and Xu, F. (1992). *Overregularization in Language Acquisition*. Monographs of the Society for Research in Child Development, Serial No. 228 (Vol. 57, No. 4). The University of Chicago Press.

Miller, G. A. (1972). English verbs of motion: A case study in semantics and lexical memory. In Melton, A. W. and Martin, E., editors, *Coding Processes in Human Memory*, chapter 14, pages 335–372. V. H. Winston and Sons, Inc., Washington, DC.

Pinker, S. (1989). *Learnability and Cognition*. The MIT Press, Cambridge, MA.

Siskind, J. M. (1994). Lexical acquisition in the presence of noise and homonymy. In *Proceedings of the Twelfth National Conference on Artifical Intelligence*, pages 760–766, Seattle, WA.

Siskind, J. M. (to appear). A computational study of lexical acquisition. Accepted for publication in *Cognition*.

Tishby, N. and Gorin, A. (1994). Algebraic learning of statistical association for language acquisition. *Computer Speech and Language*, 8(1):51–78.