A Noisy Channel Model for Systematizing Unpredictable Input Variation

Jordan J. Schneider, Laurel Perkins, and Naomi H. Feldman

1. Introduction

Children's language production is often more regular than their input. For example, after hearing an artificial language that contains unpredictable variability, children learn a grammar that is more deterministic and produces less variability than what they heard [15, 14]. Given the overwhelming evidence that children are sensitive to the statistical properties of their input while learning language (e.g. Saffran et al. [28]), it is surprising to find cases in which they do not adhere to those statistical properties. Children's regularization behavior is robust to both the amount of variability and the number of variants present in the artificial language, and it contrasts with the behavior of adults, who tend to learn a grammar that closely matches the statistics of their input [15, 14].

In this paper we focus on a case study by Singleton and Newport [29], who found similar regularization behavior in a naturalistic setting. They looked at Simon, a deaf child of non-native American Sign Language (ASL) speaking parents. At the time of the study, Simon's parents were the only speakers of ASL in his town, and so Simon likely had no access to native ASL input. Simon's parents' ASL production was significantly more variable than a control group of native ASL speakers. Despite having more variability in his input, Simon performed similarly to children of native ASL speakers on a collection of ASL production tasks.

We propose that this type of regularization behavior arises through a *noisy channel* assumption that children make during learning. Specifically, we propose that they assume some portions of their input are signal and other portions are noise with respect to the learning problem at hand. They then filter out the noise data points during the learning process and acquire grammars consistent with only the remaining signal input [24, 26]. The intuition is that if children filter out all of the low-frequency variants of a grammatical item that varies unpredictably, then they learn a grammar that disallows those variants, effectively regularizing their input. We implement a learning model that incorporates this noisy channel assumption and show that our simulated learner regularizes its input.

* Jordan Schneider, University of Texas Austin, joschnei@cs.utexas.edu, Laurel Perkins, University of Maryland, perkinsl@umd.edu, Naomi Feldman, University of Maryland, nhf@umd.edu. We thank Jeffrey Lidz, Alexander Williams, and Yevgeni Berzak for helpful discussion and advice.

 Proceedings of the 44th Boston University Conference on Language Development, ed. Megan M. Brown and Alexandra Kohut, 533-547. Somerville, MA: Cascadilla Press.

2. Noisy Channel Models

Noisy channel models assume that some of their input might be noise rather than signal generated by the source of the communication. They have to determine which parts of their input are noise that should be filtered out and ignored, and which are signal that should be used to accomplish the task at hand (in this case, language learning). Without prior knowledge about which parts of the input are noise, noisy channel models decide what is signal and what is noise by considering how likely a set of signal inputs is under a model of how the signal was generated.

A number of noisy channel models have been proposed to explain human language processing, where each sentence or segment of speech that a listener perceives is assumed to have been corrupted by a noise process. These models have captured diverse phenomena in language processing ranging from categorical perception of speech sounds [9, 18] to local coherence effects in sentence processing [20, 21] to the use of sentence fragments in conversation [3].

More recently, Perkins et al. [26] proposed a noisy channel model for explaining how children learn certain aspects of linguistic structure. The authors started from the basic observation that children may misparse portions of their input before they have acquired a fully adult-like grammar [27, 11, 23]. This means that children could internally experience a mix of well-parsed "signal" and misparsed "noise" at early stages of grammar learning. Perkins et. al. hypothesized that children might learn from this mixture of signal and noise by assuming that some proportion of their input was generated by a noise process, and filtering out that noise. They found that a noisy channel model of input filtering learned more accurately than a baseline model that did not filter its input.

If children use a noisy channel-like input filtering mechanism to filter misparses—that is, to filter noise from their *intake* [22]—it is natural to ask whether they might also use that same mechanism to filter unpredictable variability in their *input*. Here we use the model from Perkins et. al. [26] to learn from non-native input, and ask whether the model shows the type of regularization behavior that has been observed in children like Simon [29], who experience unpredictable variation in their input from non-native speakers.

3. Case Study: English Determiner Agreement

Our case study in modeling children's regularization from non-native input looks at English determiner agreement. Simon was an ASL learner and was tested on ASL grammatical morphemes, so to best simulate his learning environment, the noisy channel model would have been provided with child-directed non-native ASL. However, large ASL corpora of any kind were not available for the model to learn from, whereas corpora of texts written by late learners of English are readily available. We chose to look at English determiner agreement because non-native speakers of English often make errors in determiner agreement [16], providing a source of variation that children of non-native speakers would need to filter.

Table 1: Determiners tested in this study. Checkmarks indicate the noun types that can be used with each class of determiner. We only considered determiners that were used often enough in the input corpora, and as a result of this filtering, no determiners in our study are in the class that can be used only with mass nouns, although determiners in this class do exist (e.g. *much*).

Singular (*dog*)	Plural (*cats*)	Mass (*water*)	Class	Determiners tested
✓			*a*-class	a, an, another, each, every
	✓		*these*-class	both, these, those
✓	✓		*which*-class	which[1]
		✓	*much*-class	
✓		✓	*this*-class	this, that
	✓	✓	*all*-class	all
✓	✓	✓	*the*-class	any, no, some, the, what

In this study we considered the case of number and countability in the English determiner system, i.e., whether a particular determiner can occur with mass vs. count nouns, and singulars vs. plurals. For example, *that* can be used with only singular and mass nouns: we can say *that dog* and *that water*, but not **that cats*. Because the singular/plural distinction only occurs in the count domain, there are in principle seven possible classes of determiners that allow different patterns of combination with singular, plural, and mass nouns (Table 1). Our question is how a learner can identify, for each determiner in the input, the class that it belongs to. For the purposes of this discussion, we abstract away from questions of how number and countability are represented in the grammar, and in particular the treatment of plurals vs. mass nouns [6, 19].

4. Input Filtering Model

Our Bayesian noisy channel model assumes that its inputs are a mixture of signal and noise. It consists of a generative model, which encodes assumptions about how this mixture of signal and noise is generated, and an inference process to recover parameters of the generating process based on observed data.

Our learning model observes determiner-noun type pairs. That is, in this model we make the assumption that learners can (imperfectly) determine the number and countability of the nouns used with each determiner. When and how children identify these noun properties is a topic of debate in the literature (e.g., [5, 12, 4], but recent findings suggest that even very young learners are sensitive to the

[1] The authors had differing judgments on whether *which* can be used with mass nouns, but in this study we assume it cannot. Some approaches have argued for a typology of determiners that rules out determiners in one or more of these classes [6]; we leave for future work the question of what differences in learning such an approach would imply.

conceptual correlates of singular vs. plural and mass vs. count [8, 7, 31, 30]. It is therefore possible that children have some approximate knowledge of noun number and countability while learning determiner agreement.

The signal process uses the grammar to generate these determiner-noun type pairs. For each determiner, the grammar specifies which noun types it can be used with, and the frequency with which it occurs with each noun type. For example, *this* can be used with both singular and mass nouns, but in our corpora *this* is used much more often with singular nouns than mass nouns.

The model also assumes that some of its input was generated by a noise process, and thus might be ungrammatical. The model assumes there is a noise rate which governs what proportion of observations are generated by the noise process instead of the grammar. If an observation is generated from noise, the type of the noun is generated according to the characteristics of the noise instead of the characteristics of the grammar.

The signal and noise processes operate in parallel, generating a mixture of signal and noise as the model's input. The model does not know a priori if any particular observation was signal or noise while learning the parameters of the grammar. To resolve this, the model performs statistical inference to infer both the grammar and noise parameters simultaneously. It learns which noun types each determiner can occur with in the grammar, how often noise occurs, and with what proportion those noise instances will be each noun type.

Within the inference procedure, the model considers every partition of the observations of each determiner into signal and noise. If a partition assigns no usages of the determiner with nouns of certain types as signal (e.g., no usages of *this* with plural nouns are inferred to be signal, all are inferred to be noise), then that partition is consistent with more deterministic determiner classes (in this case, classes that disallow plural nouns). Considering all possible partitions allows the model to infer that determiners belong to more restrictive classes, even when it observes those determiners occurring in the input with all three noun types.

4.1. Generative Model

The generative model describes the learning model's assumptions about how determiner-noun type pairs in the input are generated. A specific observation of the type of a noun used with determiner d is represented by the random variable $X^{(d)}$. Each observation is drawn from either the distribution of noun types generated by the grammar or the distribution of noun types generated by the noise process.

The signal process starts with the determiner's class. For each determiner d, the set of allowed noun types is represented by $\vec{\alpha}^{(d)} = \left(\alpha_1^{(d)}, \alpha_2^{(d)}, \alpha_3^{(d)}\right)$, a three dimensional vector of binary random variables. $\alpha_1^{(d)} = 1$ if determiner d can be used with singular nouns, $\alpha_2^{(d)} = 1$ if d can be used with plural nouns, and $\alpha_3^{(d)} = 1$ if d can be used with mass nouns. The model assumes that $\vec{\alpha}^{(d)}$ is drawn

from a uniform distribution with $\frac{1}{7}$ probability for each determiner class. There is a specific distribution over noun types that governs the frequency with which determiner d is observed with each noun type represented by the random variable $\vec{\theta}^{(d)} = \left(\theta_1^{(d)}, \theta_2^{(d)}, \theta_3^{(d)}\right)$. $\vec{\theta}^{(d)}$ is drawn from a modified Dirichlet distribution with $\vec{\alpha}^{(d)}$ as its parameters. This distribution sets $\theta_i^{(d)} = 0$ when $\alpha_i^{(d)} = 0$ and otherwise samples the remaining components of $\vec{\theta}^{(d)}$ according to a Dirichlet distribution. This means that under $\vec{\theta}^{(d)}$, noun types disallowed by the determiner class will never occur, and allowed noun types can occur together with any combination of proportions. Because all non-zero $\alpha_i^{(d)} = 1$, each value of $\vec{\theta}^{(d)}$ that is consistent with $\vec{\alpha}^{(d)}$ is equally likely.

The model further assumes that each observation might have been generated by the noise process instead of the signal process. This is represented by a binary random variable e per observation that is 1 if that observation is noise and 0 otherwise. The probability of $X^{(d)}$ being generated by the noise process ($P(e = 1)$) is the value of a random variable ε, which the model assumes is drawn from the uniform distribution between 0 and 1. e is thus drawn from a Bernoulli distribution with parameter ε. When $e = 0$, that observation $X^{(d)}$ is drawn from a categorical distribution with parameter $\vec{\theta}^{(d)}$. When $e = 1$, the observation $X^{(d)}$ is instead drawn from a categorical distribution of noise with parameter $\vec{\delta} = (\delta_1, \delta_2, \delta_3)$ which has the same structure as $\vec{\theta}^{(d)}$. Our model assumes that $\vec{\delta}$ is sampled from a uniform prior distribution, $Dirichlet(1,1,1)$, so that all possible noise distributions over singular, plural, and mass nouns are equally likely a priori.

We collapse the observations $X^{(d)}$ into a vector of counts of how often each determiner was used with each noun type $\vec{k}^{(d)} = (k_1^{(d)}, k_2^{(d)}, k_3^{(d)})$. The total number of observations $n^{(d)}$ and the count of observations by noun type $\vec{k}^{(d)}$ can be split into the number and type counts of observations generated by the grammar, $n^{(d)+}$ and $\vec{k}^{(d)+}$, and number and type counts of observations generated by the noise process, $n^{(d)-}$ and $\vec{k}^{(d)-}$. In this view, the number of observations generated by the noise process $n^{(d)-}$ is sampled from a binomial distribution with parameters $n^{(d)}$ and ε, and $n^{(d)+}$ is fixed as $n^{(d)} - n^{(d)-}$. The observations generated by the grammar $\vec{k}^{(d)+}$ are then sampled from a multinomial distribution with parameters $n^{(d)+}$ and $\vec{\theta}^{(d)}$ and the observations generated by the noise process are sampled from a multinomial distribution with parameters $n^{(d)-}$ and $\vec{\delta}$.

4.2. Inference

Given the observed counts $\vec{k}^{(d)}$ for each determiner, the model uses Gibbs sampling to jointly infer the class of each determiner $\vec{\alpha}^{(d)}$, the rate of noise ε, and the characteristics of noise $\vec{\delta}$, integrating over the characteristics of signal $\vec{\theta}^{(d)}$ and the noise counts $\vec{k}^{(d)-}$.

The noise parameters ε and $\vec{\delta}$ are initialized randomly. To sample $\vec{\alpha}^{(d)}$ given ε and $\vec{\delta}$, the exact probabilities of each of the seven possible values of $\vec{\alpha}^{(d)}$ given

these noise parameters are computed for all determiners according to Bayes' rule (superscripts per determiner are omitted from here onward),

$$P(\vec{\alpha}|\vec{k},\varepsilon,\vec{\delta},n) = \frac{P(\vec{k}|\vec{\alpha},\varepsilon,\vec{\delta},n)P(\vec{\alpha}|\varepsilon,\vec{\delta},n)}{\sum_{\vec{\alpha}} P(\vec{k}|\vec{\alpha},\varepsilon,\vec{\delta},n)P(\vec{\alpha}|\varepsilon,\vec{\delta},n)} \tag{1}$$

and new values of $\vec{\alpha}$ are sampled for each determiner. After sampling all $\vec{\alpha}$, a new value of ε is sampled using ten iterations of Metropolis-Hastings sampling

$$P(\varepsilon|\vec{k},\vec{\alpha},\vec{\delta},n) \propto P(\vec{k}|\vec{\alpha},\varepsilon,\vec{\delta},n)P(\varepsilon|\vec{\alpha},\vec{\delta},n) \tag{2}$$

followed by a new value of $\vec{\delta}$ also sampled with ten iterations of Metropolis-Hastings sampling

$$P(\vec{\delta}|\vec{k},\vec{\alpha},\varepsilon,n) \propto P(\vec{k}|\vec{\alpha},\varepsilon,\vec{\delta},n)P(\vec{\delta}|\vec{\alpha},\varepsilon,n) \tag{3}$$

These sampling steps are repeated for 1000 iterations of Gibbs sampling.

The priors $P(\vec{\alpha}|\varepsilon,\vec{\delta},n), P(\varepsilon|\vec{\alpha},\vec{\delta},n)$, and $P(\vec{\delta}|\vec{\alpha},\varepsilon,n)$ are all uniform, and the likelihood $P(\vec{k}|\vec{\alpha},\varepsilon,\vec{\delta},n)$ in Equations (1) to (3) can be calculated by first considering each possible number of observations generated by the noise process n^- and then further considering each possible way to attribute n^- of the observations in $\vec{k}$ as noise and n^+ as signal

$$P(\vec{k}|\vec{\alpha},\varepsilon,\vec{\delta},n) = \sum_{n^-=0}^{n} P(n^-|\varepsilon,n) \sum_{\vec{k}^- \in K^-} P(\vec{k}^-|\vec{\delta},n^-)P(\vec{k}^+|\vec{\alpha},n^+) \tag{4}$$

where K^- is the set of values that $\vec{k}^-$ can take that are consistent with $\vec{\alpha}, n^-$, and $\vec{k}$. This likelihood is

$$P(\vec{k}|\vec{\alpha},\varepsilon,\vec{\delta},n) = \sum_{n^-=0}^{n} \frac{\binom{n}{n^-}\varepsilon^{n^-}(1-\varepsilon)^{n^+}}{\binom{n^+-1+||\vec{\alpha}||_1}{n^+}} \sum_{\vec{k}^- \in K^-} \binom{n^-}{k_1^-,k_2^-,k_3^-} \delta_1^{k_1^-}\delta_2^{k_2^-}\delta_3^{k_3^-} \tag{5}$$

5. Simulations

We conducted two simulations in which the model jointly inferred the class of each determiner, the noise rate, and the characteristics of the noise.[2] In Simulation 1, the model inferred these based on counts of how often determiners were used with each noun type in a non-native English corpus to simulate a child learning exclusively from non-native input like Simon. In Simulation 2, the model inferred the same parameters based on counts from a native English corpus, simulating children learning from native input like Simon's peers.

[2] Model inputs and code are available at https://github.com/jordan-schneider/input-filter.

5.1. Corpora and Preprocessing

Each simulation requires a corpus to model the learning environment of the children being simulated. The non-native corpus was the EFCAMDAT 2 corpus [13, 10], a corpus of text written for an online course for learning English. This corpus was chosen because it was large and had been annotated with errors, allowing us to compute the rates at which various grammatical phenomena were produced erroneously. We note that this corpus provides an imperfect model of a child's learning environment, which would ideally use a corpus of child-directed speech; however, we were unaware of a suitably large and annotated corpus of child-directed non-native English at the time of writing.

The native corpus was the ENRON corpus [17], a corpus of emails written by employees of the ENRON corporation. This corpus was chosen over other larger corpora because like the EFCAMDAT corpus, the ENRON corpus is primarily business English. Although it is not certain that all text in the ENRON corpus was written by native English speakers, there are twice as many (2.03% vs. 1.02%) ungrammatical determiner-noun pairs in the EFCAMDAT corpus compared to the ENRON corpus, and so the relative ordering between the corpora and the learning environments of the children in the Simon case study [29] is preserved.

Determiner-noun pairs were extracted from each corpus. The corpora were first cleaned by removing URLs, emails addresses, formatting marks, and other text that interfered with parsing. Next, part of speech and dependency parses of each corpus were produced using a pre-trained SyntaxNet parser [1]. From these parses, all determiner-noun pairs in each corpus were extracted. Additionally, SyntaxNet identified plural nouns. Nouns with null determiners were ignored. For each remaining noun that was not identified as plural, a noun type was assigned by referencing the CELEX2 database [2]. If that noun had definitions in CELEX as a mass noun and none as a singular, then that noun was marked as mass. All nouns that had at least one singular definition were marked as singular. Using automated methods, it was not possible to tell whether a noun that could be singular was being used as singular or mass in any given instance; this contributed an additional source of noise to the model. This issue is further compounded by the fact that nouns with only singular definitions in CELEX can sometimes be used as mass nouns; e.g. *tomato* is normally a count noun but has a mass reading in *Tomato was splattered on the wall*. This noise was present in both the native and non-native corpora.

Finally, the counts of how often each determiner was used with each noun type were aggregated, resulting in quadruplets like (*both*, 1452 singular nouns, 873 plural nouns, 262 mass nouns). Any determiner whose total count in either corpus was under 500 instances was omitted. These counts were then downsampled by taking the logarithm of each count, then normalizing all log-counts so the total input size was 5000 determiner-noun type pairs. The log-weight was used to compensate for the unbalanced distribution of determiners; specifically, *the* comprised 49% of all determiners in the EFCAMDAT corpus and 55% of all determiners in the ENRON corpus.

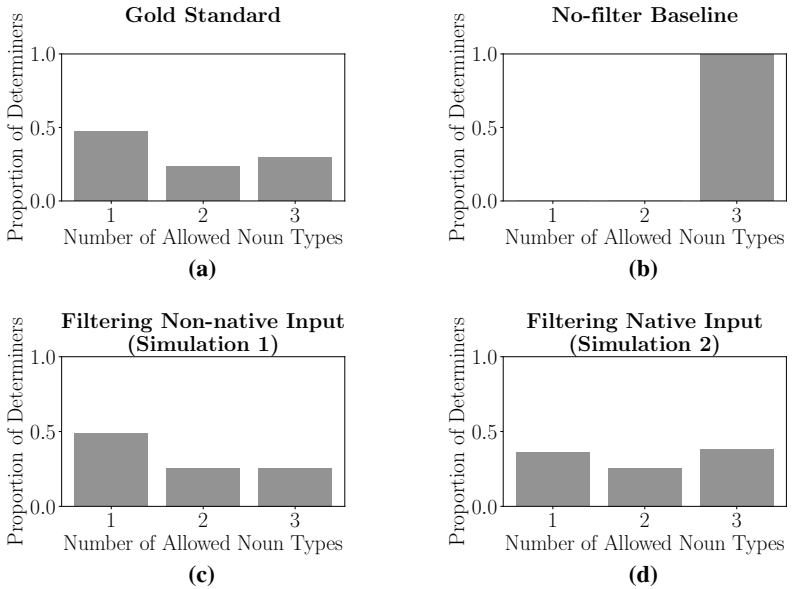

Figure 1: How many determiners allow one, two, or three noun types under the gold standard, and under the determiner class inferred by a non-filtering baseline and by the filtering model trained on the non-native corpus and the native corpora.

5.2. Simulation 1: Learning from Non-native English

The primary goal of this study is to suggest an input filtering mechanism to model regularization. We test if our model can regularize by comparing its output to that of a non-filtering baseline, and to a set of gold-standard determiner classes specified by the authors (as given in Table 1). Specifically, we ask how often the inferred determiner classes allow one, two, or all three noun types. For example, *the* allows all three types, *this* allows two (singular and mass), and *a* allows one (singular). The raw output of the model is a probability that each determiner is in each of the seven determiner classes, but for this analysis we say that a determiner is in the class with the highest probability under our model. The fewer noun types allowed by the class, the fewer noun types will be produced after learning, and so the more regular the language production will be under the acquired class.

Figure 1 shows how often our model and the non-filtering baseline inferred determiner classes that allow each number of noun types, as well as the number of noun types in the gold-standard. Larger bars for the classes that allow only one or two noun types indicate more deterministic acquired classes, in aggregate. The non-filtering baseline (Figure 1b) observes all determiners with all noun types, does not filter any of those observations, and so infers that all determiners can be used with all noun types. Since our model (Figure 1c) and the gold-standard (Figure 1a) both assign some determiners to classes that allow one or two noun types, they are both more deterministic than the non-filtering baseline. Our model is slightly more

deterministic than the gold-standard since it infers that more determiners can be used with both one (49% vs. 47%) and two (25% vs. 24%) noun types, and fewer determiners can be used with all three noun types (26% vs. 29%). This means that, in aggregate, our model not only regularizes, but slightly over-regularizes.

We are also interested in determining if our model succeeds in its job as a noisy channel model: does it separate signal from noise? We can answer this question by comparing the determiner classes inferred by our model to the gold-standard determiner classes. If the model assigns the highest probability to the gold-standard determiner class, then it filtered the disallowed noun types as noise, and did not filter the allowed noun types. We compare the accuracy of our model relative to this gold-standard to the accuracy of a learner that guessed each determiner class uniformly, and the accuracy of the non-filtering baseline. Figure 2 shows that our model acquired the gold-standard class for ten of the 17 determiners (59%). This is well above the accuracy of both the random baseline of one in seven correct (14%), and the non-filtering baseline, which infers that all determiners can be used with all three noun types, yielding five in 17 correct (29%). This indicates that the model is effectively separating signal from noise.

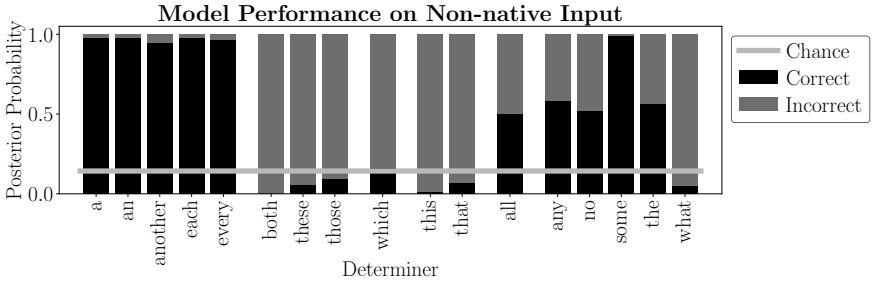

Figure 2: The posterior probability assigned to the correct and incorrect determiner classes for each determiner in Simulation 1. On the x-axis are all the determiners in the study grouped by the gold-standard determiner class. Black bars represent the posterior probability the model assigned to the correct class for that determiner, and dark grey bars represent the sum of the probabilities the model assigned to the incorrect determiner classes. A light grey horizontal line marks the chance probability of one in seven that guessing uniformly at random would achieve.

Figure 2 aggregates the probability assigned to all incorrect classes, but looking at the probability of each category individually can provide insight into why the model regularized and what kind of mistakes the model made. Figure 3 show these probabilities for *the*, *which*, and *both*. In Figure 3a, the model assigns 56% probability to the correct determiner classes for *the*, which allows all three noun types. However, in other cases, the model made mistakes by over- or under-regularizing its input. For example, *that* can occur with both singular and mass nouns. However, Figure 3b shows that the model assigned the most probability to the class that allows only singular nouns, implying that it interpreted the mass

usages of *that* as noise. The incorrect determiner class that the model learned was more deterministic than the gold-standard determiner class. Figure 3c shows that the model made the opposite mistake with *both*. *Both* can only occur with plural nouns, but the model assigned a 94% probability that *both* can occur with all noun types. In this case, the model failed to filter out the noise singular and mass nouns and so learned a less deterministic determiner class than the gold-standard. Of the seven determiners for which the model disagreed with the gold standard, three are under-regularized and four are over-regularized.

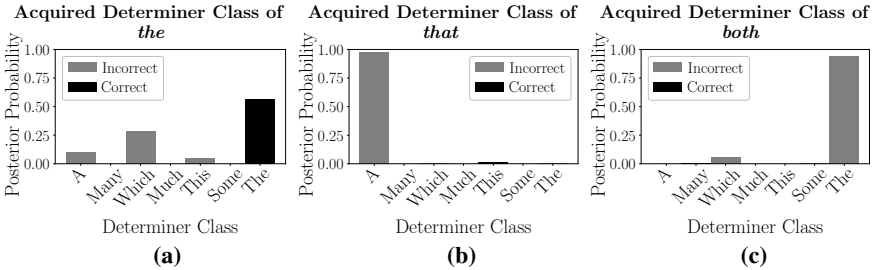

Figure 3: The probability assigned by the model to each determiner class for *the*, *that*, and *both*. On the x-axis are the seven possible determiner classes, labeled by an example determiner in that class. On the y-axis is the posterior probability the model assigned to each of the determiner classes after training on the non-native dataset. The correct class is in black.

The model's inferred noise parameters are difficult to interpret but are presented for later comparison. The mean noise rate sampled was 15%, and the mean noise characteristics would generate noise as 47% singular, 26% plural, and 26% mass nouns. Although the corpus has only 2% ungrammatical determiners, we cannot say that the 15% overall noise rate is inappropriate as the model's noise rate includes observations generated by the noise process that happen to be grammatical by chance. Additionally, it is plausible that the risk of being misled by noise is greater than the benefit to learning from having additional input to learn from, causing children to prefer to over-filter early in learning. Further study is necessary to evaluate the appropriateness of these noise parameters.

Our results show that our noisy channel model regularized its input and learned accurate determiner classes compared to a baseline, behaving qualitatively like Simon in the Singleton & Newport case study [29]. We cannot directly compare our results to Simon's behavior because the learning environments and tasks were not matched. But qualitatively, Simon and our model were both given variable input and, relative to that input, produced more regular language and a more regular grammar respectively. Overall, our first simulation provides evidence that noisy channel models regularize similarly to children learning from non-native input.

Although we have demonstrated that our model can regularize like Simon, we have yet to show that our model is similar to Simon in performing just as well as

the children of native speakers. To test this we ran a second simulation in which our model was given native English input. If our model were like Simon, then it would perform similarly well after learning from both native and non-native input, suggesting that an input filtering mechanism could fully explain Simon's performance. If our model performed differently on native input and non-native input, then it would be unlike Simon in this regard. This may indicate that Simon is using another mechanism in addition to an input filter, or that our simulation does not accurately imitate Simon's learning environment.

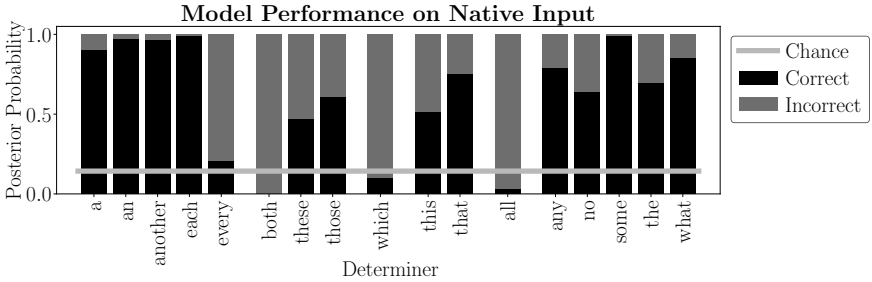

Figure 4: The model learns the gold-standard determiner classes of *those*, *this*, *that*, and *what* in addition to those in Simulation 1. The model did not learn the gold-standard properties for *every* and *all* despite doing so in Simulation 1. Overall the model is more accurate when learning from native input.

5.3. Simulation 2: Learning from Native English

We evaluate if our model performed similarly when learning from native and non-native input using the regularity and accuracy measures from Simulation 1. We found that our model regularized its input to a lesser extent when learning from native English input than when learning from non-native input, and acquired more accurate determiner classes overall. This indicates that our model is not like Simon in learning a native-like grammar from non-native input.

Using the same regularity metric as in Section 5.2, Figure 1 shows that in Simulation 2 the model (Figure 1d) regularized its input much more than a non-filtering baseline (Figure 1b) again because the model infers that many determiners can be used with only one or two noun types. However, the model regularized less than it did in Simulation 1 (Figure 1c), inferring that fewer determiners can be used with one noun type (36% vs. 49%), roughly the same can be used with two noun types (26% vs. 25%), and more can be used with all three noun types (38% vs. 26%). The noisy channel model regularizes regardless of its input, but when there is less noise to filter, the model regularizes less.

This difference in regularization comes with an increase in the accuracy of the acquired determiner classes. In Simulation 2, the model acquired the gold-standard determiner class for 12 of 17 (70%) of determiners, well above both the chance and non-filtering baselines of one in seven (14%) and five in 17 (29%), respectively.

Comparing Figure 2 and Figure 4 shows that in Simulation 2, the model learns the gold-standard determiner classes for *those*, *this*, *that*, and *what* in addition to all determiners correctly learned in Simulation 1 except *every* and *all*. *This*, *that*, and *what* all allow two or three noun types and were over-regularized by the model in Simulation 1.

Where the model had a mix of under- and over-regularization mistakes in Simulation 1, all five mistakes in Simulation 2 were due to under-regularization. This can be explained as a consequence of the different amounts of noise present in the native vs. non-native corpora. The model correctly learns that there is less noise in the native corpus than the non-native corpus, inferring a distribution over the noise rate with lower mean (12% in Simulation 2 corpus vs. 15% in Simulation 1) while the characteristics of the noise did not change much (56% singular, 18% plural, and 25% mass in Simulation 2 vs. 47% singular, 26% plural, and 26% mass in Simulation 1). This lower noise rate in turn increases the probability that input was generated from the grammar and so the generative model assigns higher probability to less deterministic determiner classes, which can generate these examples through the signal process. In aggregate, the lower noise rate caused the model to regularize less in Simulation 2 than in Simulation 1.

As our model regularized less and was more accurate when learning from native input than from non-native input, it does not replicate Simon's behavior of acquiring a native-like grammar. We consider various reasons for this difference in the discussion below.

6. Discussion

We tested whether a noisy channel model could account for the regularization behavior seen in child language learners by applying the model to the problem of learning the agreement behavior of determiners in English. Specifically, we tested if a noisy channel model adapted from Perkins et al. [26] would regularize input from non-native speakers of English, paralleling the behavior of a child of non-native speakers of ASL in Singleton & Newport [29]. We found that, despite all determiners occurring with all noun types in the input, the model acquired more deterministic classes for 74% of determiners, demonstrating that noisy channel models like ours can account for regularization.

Furthermore, the model acquired the gold standard determiner classes for ten of the 17 determiners in its input, despite not being optimized for the accuracy of its acquired determiner classes and never receiving feedback about its accuracy. In addition to being able to account for child regularization behavior, the noisy channel model also solves a variable input problem in an effective way so that accurate learning can occur in the presence of statistical noise.

However, the model did not perfectly reproduce the findings of Singleton & Newport [29]. In their case study, Simon's production was similar to that of children of native speakers, and so an accurate model would have acquired similar determiner classes from native and non-native input. Our model learned the correct

classes for two more determiners when learning from the native input than it did learning from the non-native input, a 12% difference in accuracy. Where errors were made in both simulations, the model also made different mistakes, both over- and under-regularizating on non-native input but always under-regularizing on native input. These differences between the model's output on native and non-native input are the largest way our model does not match Simon's performance.

There are many possible reasons for this difference between our model and Simon's performance. It may be a result of using a corpus of non-child directed speech to simulate a child's learning. Alternatively, acquisition of determiner agreement from non-native English may not be parallel to the acquisition of verb morphology from non-native ASL; assessing this possibility would require additional empirical data. Finally, it is also possible that learners like Simon need additional mechanisms beyond a noisy channel assumption to acquire language from variable input [24]. Further research is needed to determine which, if any of these, are responsible for the difference between the model and the observations.

In Bayesian models, a noisy channel allows a model to consider parameter values that are in the hypothesis space, but are inconsistent with its observations. In doing so, the noisy channel reinforces prior biases by allowing the model to attribute observations that are unlikely under the prior to noise. It also makes the bias more robust: whereas a bias encoded in the prior is eventually overwhelmed by evidence in most Bayesian models, a noisy channel model can continue to account for some proportion of its data as noise, regardless of the absolute quantity of data.

Noisy channel models can also provide an account of how, over time, adults come to match their input instead of regularizing. As adults understand more and more of the language that they hear, noise from misparses is eliminated. Adults might eventually come to believe that much less of their input is noise than children do (i.e. they infer that the value of ε is lower). This would lead them to believe that more new language information that they hear is signal, and so become more perfect statistical learners. In the limit, if adults using an input filtering model believe there is no noise, then they would match the statics of their input exactly.

The similarity between the model presented in this paper and the models proposed by Perkins et. al. [26, 25, 24] suggests that there might be a general input filtering mechanism that children use while learning language, and that regularization occurs because of this more general process. There are many sources of statistical noise which might be present in children's input or intake before they have acquired the grammatical knowledge to represent their input veridically. Thus, there are many circumstances in which children might benefit from filtering their input while learning. If a noisy channel-like mechanism is filtering these sources of noise, that same mechanism might filter noise from unpredictable variation and so be the cause of regularization.

References

[1] Daniel Andor et al. “Globally Normalized Transition-Based Neural Networks”. In: *arXiv:1603.06042 [cs]* (June 8, 2016).

[2] R H. Baayen, R Piepenbrock, and L Gulikers. *CELEX2 LDC96L14*. 1995.

[3] Leon Bergen and Noah D. Goodman. “The Strategic Use of Noise in Pragmatic Reasoning”. In: *Topics in Cognitive Science* 7 (2015), pp. 336–350.

[4] Paul Bloom. “Semantic competence as an explanation for some transitions in language development”. In: *Other children, Other languages–Theoretical issues in language acquisition. Hillsdale, NJ: Erlbaum* (In press).

[5] Roger W. Brown. “Linguistic determinism and the part of speech.” In: *The Journal of Abnormal and Social Psychology* 55.1 (1957), p. 1.

[6] Gennaro Chierchia. “Plurality of mass nouns and the notion of “semantic parameter””. In: *Events and grammar*. Springer, 1998, pp. 53–103.

[7] Lisa Feigenson and Susan Carey. “Tracking individuals via object-files: evidence from infants’ manual search”. In: *Developmental Science* 6.5 (2003), pp. 568–584.

[8] Lisa Feigenson, Susan Carey, and Marc Hauser. “The representations underlying infants’ choice of more: Object files versus analog magnitudes”. In: *Psychological Science* 13.2 (2002), pp. 150–156.

[9] Naomi H. Feldman, Thomas L. Griffiths, and James L. Morgan. “The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference”. In: *Psychological Review* 116.4 (2009), pp. 752–782.

[10] Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, et al. “Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT)”. In: *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*. 2013, pp. 240–254.

[11] Lila Gleitman. “The Structural Sources of Verb Meanings”. In: *Language Acquisition* 1.1 (1990), pp. 3–55.

[12] Peter Gordon. “Evaluating the semantic categories hypothesis: The case of the count/mass distinction”. In: *Cognition* 20.3 (1985), pp. 209–242.

[13] Yan Huang et al. “Dependency parsing of learner English”. In: *International Journal of Corpus Linguistics* 23.1 (2018), pp. 28–54.

[14] Carla L. Hudson Kam and Elissa L. Newport. “Getting it right by getting it wrong: When learners change languages”. In: *Cognitive Psychology* 59.1 (2009), pp. 30–66.

[15] Carla L. Hudson Kam and Elissa L. Newport. “Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change”. In: *Language Learning and Development* 1 (Apr. 2005), pp. 151–195.

[16] Jacqueline S. Johnson and Elissa L. Newport. “Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language”. In: *Cognitive Psychology* 21.1 (Jan. 1, 1989), pp. 60–99.

[17] Bryan Klimt and Yiming Yang. “The Enron Corpus: A New Dataset for Email Classification Research”. In: *Machine Learning: ECML 2004*. Ed. by Jean-François Boulicaut et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, pp. 217–226.

[18] Yakov Kronrod, Emily Coppess, and Naomi H. Feldman. “A unified account of categorical effects in phonetic perception”. In: *Psychonomic Bulletin and Review* 23.6 (2016), pp. 1681–1712.

[19] Peter Lasersohn. “Mass Nouns and Plurals”. In: *Semantics: An International Handbook of Natural Language Meaning*. Ed. by Claudia Maienborn, Klaus von Heusinger, and Paul Portner. De Gruyter Mouton, 2011, p. 2.

[20] Roger Levy. “A noisy-channel model of rational human sentence comprehension under uncertain input”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2008).

[21] Roger Levy et al. “Eye movement evidence that readers maintain and act on uncertainty about past linguistic input”. In: *Proceedings of the National Academy of Sciences* 106.50 (2009), pp. 21086–21090.

[22] Jeffrey Lidz and Annie Gagliardi. “How Nature Meets Nurture: Universal Grammar and Statistical Learning”. In: *Annual Review of Linguistics* 1.1 (2015), pp. 333–353.

[23] Jeffrey Lidz and Lila Gleitman. “Yes, we still need Universal Grammar”. In: *Cognition* 94 (Nov. 2004), pp. 85–93.

[24] Laurel Perkins. “How Grammars Grow: Argument Structure and the Acquisition of Non-Basic Syntax”. PhD thesis. 2019.

[25] Laurel Perkins, Naomi H. Feldman, and Jeffery Lidz. “Mind the Gap: Learning the Surface Forms of Movement”. Boston University Conference on Language Development. 2019.

[26] Laurel Perkins, Naomi Feldman, and Jeffrey Lidz. “Learning an input filter for argument structure acquisition”. In: *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*. 2017, pp. 11–19.

[27] Steven Pinker. *Language Learnability and Language Development (1984/1996)*. Cambridge, MA: Harvard University Press, 1996.

[28] Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. “Statistical Learning by 8-Month-Old Infants”. In: *Science* 274.5294 (1996), pp. 1926–1928.

[29] Jenny L. Singleton and Elissa L. Newport. “When learners surpass their models: The acquisition of American Sign Language from inconsistent input”. In: *Cognitive Psychology* 49.4 (2004), pp. 370–407.

[30] Nancy N. Soja, Susan Carey, and Elizabeth S. Spelke. “Ontological categories guide young children’s inductions of word meaning: Object terms and substance terms”. In: *Cognition* 38.2 (1991), pp. 179–211.

[31] Elizabeth S. Spelke. “Perception of Unity, Persistence, and Identity: Thoughts on Infants’ Conceptions of Objects”. In: *Neonate Cognition: Beyond the Blooming Buzzing Confusion*. Ed. by Jacques Mehler and R. Fox. Lawrence Erlbaum, 1985, pp. 89–113.

Proceedings of the 44th annual Boston University Conference on Language Development

edited by Megan M. Brown
and Alexandra Kohut

Cascadilla Press Somerville, MA 2020

ISSN 1080-692X
ISBN 978-1-57473-057-9 (2 volume set, paperback)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com