Optimal Language Learning: The Importance of Starting Representative

Anna N. Rafferty (rafferty@cs.berkeley.edu)
Computer Science Division, University of California, Berkeley, CA 94720 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

Child-directed speech has a distinctive structure and may have facilitatory effects on children's language learning. We consider these facilitatory effects from the perspective of Marr's levels of analysis: could they arise at the computational level or must they be located at at the algorithmic or implementation levels? To determine if the effects could be due to computational level benefits, we examine the question of what samples from a language should best facilitate learning by identifying the optimal linguistic input for an ideal Bayesian learner. Our analysis leads to a mathematical definition of the "representativeness" of linguistic data, which can be computed for any probabilistic model of language learning. We use this measure to re-examine the debate over whether language learning can be improved by "starting small" (i.e. learning from data that have limited complexity). We compare the representativeness of corpora with differing levels of complexity, showing that while optimal corpora for a complex language are also complex, it is possible to construct relatively good corpora with limited complexity. We discuss the implications of these results for the level of analysis at which a benefit of starting small must be located.
Keywords: language learning; child-directed speech; Bayesian models; representativeness; starting small

Introduction

Child-directed speech is an important source of information for children's language acquisition. Hoff and Naigles (2002) found that the amount of child-directed speech produced by mothers was predictive of the vocabulary of their children, and Cameron-Faulkner, Lieven, and Tomasello (2003) found correlations between the grammatical frames mothers used in speech to their children and the grammatical frames used by the children. Child-directed speech also differs from adult-directed speech in a number of ways. For example, Snow (1972) found that speech to two year olds by caregivers has simplified structure and involves more repetitions than speech to older children or adults, and Sherrod, Friedman, Crawley, Drake, and Devieux (1977) found that the mean length of utterances spoken to a child changed in response to changes in the child's understanding. Overall, child-directed speech tends to be simplified, more grammatically correct, and more repetitive than adult-directed speech (Pine, 1994). This raises an important question: Does the structure of child-directed speech facilitate language acquisition?

There is some evidence for a facilitatory effect of child-directed speech. Furrow, Nelson, and Benedict (1979) found that children's language development was positively correlated with mothers' use of simple constructions, and Newport, Gleitman, and Gleitman (1977) found that acquisition of certain syntactic features was facilitated by characteristics of mothers' speech, such as placement of particular syntactic structures early in sentences. However, Newport et al. (1977) also found that many measures of acquisition were unaffected by characteristics of caregivers' speech, and Huttenlocher, Vasilyeva, Cymerman, and Levine (2002) found that exposing children to more complex speech resulted in the children using more complex syntax.

Previous work has used specific computational models such as associative learning and artificial neural networks to explore the effects of simplified input on language learning (Goldowsky & Newport, 1993; Elman, 1993; Rohde & Plaut, 1999). Elman (1993) found that training a simple recurrent neural network to predict the next word in a sequence using a corpus of limited complexity resulted in better generalization than beginning with the full corpus. However, the effects of "starting small" are far from clear: Rohde and Plaut (1999) subsequently found a disadvantage for language learning that begins with data of limited complexity when using similar models and corpora.

Demonstrating an effect of starting small under specific assumptions about learning leaves open the question of the level of analysis at which there might be an advantage for child-directed speech. Marr (1982) defined three levels at which information processing systems can be analyzed: the *computational* level, where the analysis aims to identify the abstract problem being solved and its ideal solution; the *algorithmic* level, where the focus is on the representation and algorithm being used to implement this solution; and the *implementation* level, which emphasizes the physical hardware on which the algorithm is executed. Facilitatory effects of the structure of child-directed speech could be caused by considerations at any of these levels. At the computational level, data of this kind could provide more statistical evidence for the structure of the language. Alternatively, constraints at the algorithmic or implementation levels might limit the information-processing capacities of children, making simplified input necessary despite the lack of a computational level benefit.

We try to identify the level of analysis at which a benefit from simplified input could be located by asking what characteristics a sample of language should have in order to be most useful for an ideal learner. If simpler corpora are better for this ideal learner, then we can provide a computational-level account of the benefit of starting small. If not, such an effect must be located at a lower level. It is necessary to consider the performance of ideal learners in order to rule out the possibility that starting small provides a computational-level advantage. If this were the case, it would not be necessary to assume algorithmic level constraints are the cause of an

advantage for starting small, as has been done in previous research.

We identify the optimal input for an ideal Bayesian language learner by asking what data maximize the posterior probability such a learner ascribes to the target language. This is a special case of the problem of defining a "representative" sample analyzed by Tenenbaum and Griffiths (2001). Consequently, we define a Bayesian measure of representativeness, and use this measure to give a mathematical characterization of an optimal corpus. We present a general mathematical result characterizing representativeness for discrete probability distributions, which are the basic component of any probabilistic model of language. This result provides the basis for a more detailed exploration of whether language of limited complexity might be as good or better for learning than language of full complexity. We explore the implications of this result by identifying the optimal input for four different learning scenarios, involving estimating probabilistic grammars with varying degrees of knowledge about the structure of a language and estimating n-gram models.

Identifying Optimal Corpora

To understand the characteristics of an optimal sample of language, we formalize the problem of language learning in terms of Bayesian inference. Learning a probabilistic model of language requires estimating the value of a set of parameters θ from observed linguistic data d. Assuming the learner has some initial beliefs about the value of θ, expressed through a *prior* probability distribution $p(\theta)$, the beliefs of a rational learner after observing d are given by the *posterior* distribution $p(\theta|d)$ obtained by applying Bayes' rule,

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{\int p(d|\theta)p(\theta)d\theta} \quad (1)$$

where the *likelihood* $p(d|\theta)$ indicates the probability of d under the probabilistic model with parameters θ.

A Measure of Representativeness

Formalizing language learning in this way now allows us to consider what corpora will most strongly support learning. Assume that the true structure of the language is characterized by parameters θ^*; we consider a learner that is simply learning this structure, although more complicated models that also learn other parts of the language, such as semantics, are possible. To maximize the probability of a learner inferring θ^* over other values of θ, a teacher should provide data d that maximize $p(\theta^*|d)$. Examination of the right hand side of Equation 1 shows that this can be done by maximizing

$$R(d,\theta^*) = \frac{p(d|\theta^*)}{\int p(d|\theta)p(\theta)d\theta} \quad (2)$$

with respect to d, as the prior probability $p(\theta^*)$ is constant and thus unaffected by the choice of d. Tenenbaum and Griffiths (2001) suggested that $R(d,\theta^*)$ be considered a measure of the "representativeness" of d relative to θ^*, being an indicator of the strength of evidence that d provides in favor of θ^* relative to other values of θ. Intuitively, a sample is more representative if it is both very probable under the true model (the numerator of Equation 2) and not as probable under a model selected at random (the denominator of Equation 2).

Representativeness for Discrete Distributions

In general, we may not be able to solve the integral in the denominator of Equation 2 exactly. However, we can solve this integral in the case where the model $p(d|\theta)$ is a discrete probability distribution, as is often true with probabilistic models of language. For a multinomial with ordered outcomes, the likelihood is $p(d|\theta) = \prod_{i=1}^{t}(\theta_i^*)^{k_i}$, where t is the number of possible outcomes, θ_i^* is the probability of outcome i, and k_i is the number of times the outcome i occurred. We place a uniform Dirichlet prior on the distribution θ, reflecting no strong expectations about the probabilities of different rules. Thus, the integral in Equation 2 is in this case:

$$\int_\Delta \prod_{i=1}^{t} \theta_i^{k_i} d\theta = \frac{(\prod_{i=1}^{t} k_i!)}{(t-1+\sum_{i=1}^{t} k_i)!} = \frac{(\prod_{i=1}^{t} k_i!)}{(t-1+n)!} \quad (3)$$

where Δ is the simplex of values such that $\sum_{i=1}^{t}\theta_i = 1$, and n is the total number of observations. The representativeness of a corpus with respect to this model with a particular value of θ is then:

$$R(d,\theta) = \frac{(t-1+n)!\prod_{i=1}^{t}\theta_i^{k_i}}{(\prod_{i=1}^{t} k_i!)}. \quad (4)$$

The optimal corpus is that which maximizes this quantity.

We can find an exact expression for the frequencies an optimal corpus would have by maximizing the quantity in Equation 4 with respect to k_i. Since the logarithm is monotonic, the corpus that maximizes $R(d,\theta)$ is also the corpus that maximizes $\log R(d,\theta)$, so we perform our maximization with this transform. Additionally, this is a constrained optimized problem since n must equal $\sum_{i=1}^{t} k_i$. We enforce this constraint with a Lagrange multiplier, and replace the factorials using Stirling's approximation to obtain the objective function:

$$L = (t-1+n)\log(t-1+n) + 1 - t + \sum_{i=1}^{t} k_i\log(\theta_i) - k_i\log(k_i) + \lambda(n - \sum_{i=1}^{t} k_i) \quad (5)$$

where λ is the Lagrange multiplier. To determine the optimum of this objective function, we differentiate with respect to k_i, set the derivative to zero, and solve for k_i. This shows that the optimal value is $k_i = n\theta_i$. Rounding to the nearest integer, this corresponds to what one might intuitively expect: The most representative corpus is that in which the relative frequencies of the outcomes match their probabilities under the target multinomial.

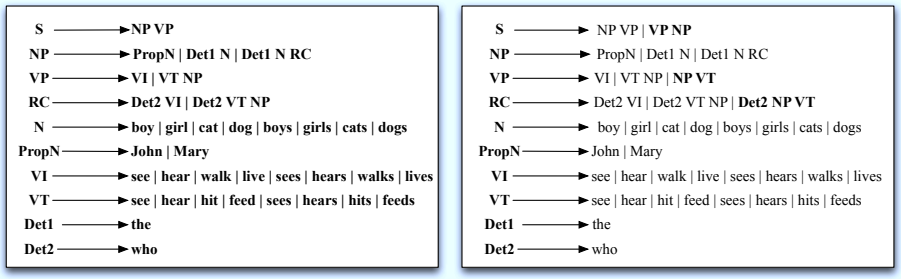

Figure 1: The context-free grammars used in our simulations. On the left, the true grammar; on the right, the overly general grammar with added rules, used in the third simulation. Bolded expansions are those present in the expanded grammar but not in the true grammar. In addition to these rules, subject-verb agreement is enforced, resulting in a much larger PCFG.

Representativeness for Probabilistic Grammars

The results for the representativeness of samples from multinomial distributions can be used to characterize optimal corpora for any probabilistic language model with discrete elements, such as an n-gram model. These results also generalize naturally to a representativeness measure for more sophisticated probabilistic models of language, such as probabilistic grammars. A *probabilistic context-free grammar* (PCFG; Baker, 1979) defines a probability distribution over sentences via a set of expansion rules for non-terminals (e.g. a noun phrase consists of a determiner followed by a noun) and distributions over those rules indicating the probability of a given non-terminal being expanded to a particular sequence (see Figure 1). The distributions over rules are independent multinomials, allowing us to build on the representativeness analysis above. In this case, the parameters θ describe the multinomial distributions associated with each expansion rule.

When the structure of sentences (i.e. the sequence of expansion rules used in generating each sentence) is known, the representativeness of a corpus follows directly from our result for multinomials. Since each rule is associated with an independent multinomial, the representativeness is the product of the representativeness for each multinomial. Thus, a representative corpus is one in which the relative frequencies with which expansion rules are used match the probabilities associated with those expansion rules in the grammar.

When the structure of sentences is unknown, $p(d|\theta)$ is obtained by marginalizing over possible structures. For PCFGs, this can be done efficiently using a dynamic program; in our simulations, we used Mark Johnson's implementation of this algorithm.[1] However, since there is not a closed form for this marginalization, we cannot calculate the denominator of Equation 2 exactly. In this case, we can use a Monte Carlo method to approximate the integral and obtain an estimate of the representativeness of a corpus.

Starting Small

As described in Elman (1993), "starting small" involves showing a learner only a limited number of "complex" sentences from a language first, and gradually exposing the learner to the full language. A sentence is complex if it contains a recursive rule; for example, in both Elman (1993) and Rohde and Plaut (1999), complex sentences are those that contain relative clauses. Both Elman (1993) and Rohde and Plaut (1999) used neural networks that learned to predict the next word in the sentence. Elman (1993) found that starting small was essential for his model; when a corpus of full complexity was used, the learner was never able to predict the next word with satisfactory accuracy. Rohde and Plaut (1999) found, in contrast, that in most cases starting small resulted negative impacts on performance, and none of their simulations showed any advantage to starting small.

We use the analysis of representativeness for an ideal language learner given in the previous section to explore the locus of a potential effect of starting small. Since our analysis focuses solely on the statistical evidence a corpus provides in favor of a particular language, we can examine whether a potential benefit of starting small could arise at the computational level, or must be a consequence of specific information-processing constraints associated with human learning. Thus, we consider two questions: First, does starting small result in particularly good corpora for language learning? And second, can a corpus of limited complexity be as good as a corpus without limited complexity? Clearly, if a starting small corpus is optimal, then such a corpus is as or more representative than a more complex corpus. However, even if a limited complexity corpus is non-optimal, it might be as representative as corpora generated by other means. In particular, we compare corpora of different complexity generated by maximizing representativeness and generated randomly.

As in the analysis in the previous section, we consider two types of corpora: those in which sentence structure is known and those in which structure is unknown. In two simulations, we assume that the learner knows the rules of the grammar, but does not know the frequencies with which they occur. Our third simulation introduces ambiguity about the rules of the grammar, and the fourth considers the possibility that children are not learning a grammar but simply distributions over which words follow one another. We used a PCFG similar to that in Elman (1991). The only instance of recursion was in the relative clause, which occurred in 75% of sentences generated from the grammar, and the grammar enforced the

[1] Version last updated 2 September 20007, and available at http://www.cog.brown.edu/∼mj/Software.htm

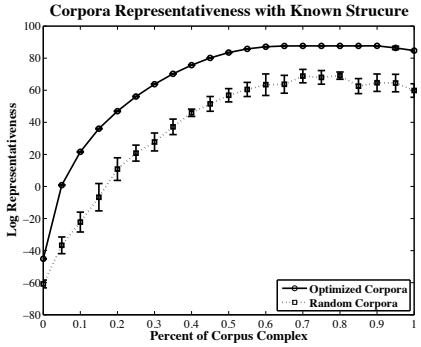

Figure 2: Representativeness of corpora with known structure. As the number of complex sentences increases, the representativeness of the corpora increases non-linearly.

agreement of subjects and verbs[2]. Figure 1 shows the grammar prior to integrating the constraint of verbal agreement; the final grammar consisted of 63 rules and 23 nonterminals.

Representativeness with Known Structure

We first considered the problem of learning from a corpus in which the structures of the sentences are known, allowing us to use the closed form given in Equation (4) to exactly compute the representativeness of the corpus. We sought to quantify how representative a corpus could be given the constraint on complexity and discover how this compared to randomly generated corpora as well as more complex corpora.

To investigate this question, we generated several types of corpora. All corpora were created by choosing a subset of sentences from a large corpus generated by the grammar. *Random* corpora were generated by selecting this subset randomly, subject to a constraint on the number of complex sentences. *Optimized* corpora were collections of sentences chosen to maximize representativeness. An ε-greedy perturbation process was used to maximize representativeness. First, an initial corpus of the target complexity was randomly selected. This corpus was perturbed by adding additional sentences, and then pruning sentences from the augmented corpus. With small probability, a sentence was chosen randomly to add or prune. Otherwise, a sentence was chosen by checking the effect of adding or pruning each possible sentence and greedily adding or pruning the sentence that resulted in the corpus with the largest representativeness. Twenty perturbations of ten sentences each were performed; results were not sensitive to small variations in these parameters.

For both the random and optimized conditions, we created corpora with constrained complexity. Corpora were generated with complexity ranging from 0% to 100% complex sentences, at 5% intervals. A complex sentence was any sentence containing a relative clause. Additionally, random and optimized corpora were generated with no complexity constraint. Each corpus contained 100 sentences.

As shown in Figure 2, this procedure succeeds in finding subsets of sentences that are significantly more representative than randomly generated corpora of the same complexity. However, the limitation on complexity greatly affects representativeness. While limiting complexity significantly impacts the representativeness of only one rule, that which allows the introduction of the relative clause, this impact is severe enough to outweigh the representativeness of the other rules. Thus, an optimized sample of severely limited complexity is much less representative than a random sample in which complexity is not constrained. When the limit is not as severe, though, optimized corpora with somewhat limited complexity and random corpora with greater complexity have equal representativeness, due to the fact that the severity of the complexity constraint has a non-linear effect on representativeness (Figure 2). For corpora of unconstrained complexity, the results mirror the results for corpora with constrained complexity equal to the true base rate of complex sentences for the grammar (75%). The average representativeness of randomly selected corpora was 65.7 ± 4.9, with $76.3\% \pm 4.9$ complex trees, while the average representativeness of corpora selected for representativeness was $87.7 \pm 8x10^{-5}$, with $80.1\% \pm 10.3$ complex trees.

Representativeness with Unknown Structure

The previous simulation assumed that our corpus consisted of the structure of the sentences, from which we could directly compute the representativeness of a given corpus. However, one might alternatively assume that a language learner has only the sentences as data and must consider all possible structures. We examine this possibility by using the same corpora of sentences as in the previous simulation, but assuming the structure of each sentence is unknown.

As mentioned in the previous section, when the structure of sentences is unknown we need to resort to Monte Carlo approximation to compute representativeness. We used importance sampling (Neal, 1993); our proposal distribution was a Dirichlet distribution with parameters equal to the true distribution in the grammar multiplied by ten. Results are averages of 30 iterations of 10,000 samples each; in the case of random corpora, sampling was done for each of ten corpora with the same constraints on complexity. Given that variance for the optimized corpora was much smaller, sampling was done for only one optimized corpus of each level of complexity. We consider the same four levels of complexity as Elman (1993) and Rohde and Plaut (1999): 0%, 25%, 50%, or 75% of the total corpus size. Additionally, we consider corpora of unconstrained complexity.

Figure 3 shows that the general trends from the previous simulation hold, with a few variations. The separation between the optimized corpora and the random corpora is smaller than when the structure is known. This is partially due to the way the corpora were created. Presumably, if it was feasible to optimize over corpora with unknown structure, further separation might be attained. However, these results do suggest that optimizing the input sentences would not greatly help a learner who must consider all possible structures of sentences. Consistent with the previous simulations, repre-

[2]Grammar creation was facilitated by the Simple Language Generator (Rohde, 2003)

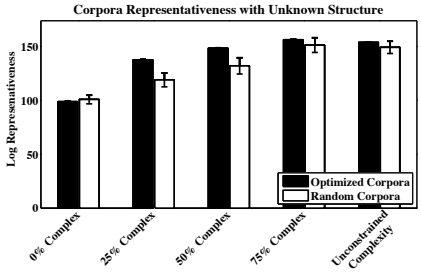

Figure 3: Representativeness of corpora with unknown structure. Limiting the complexity of a corpus limits its representativeness, with the most extreme limitation having the greatest effect.

sentativeness increases non-linearly with complexity. Again, the least complex corpora are not as representative as those that match the base rate of complexity in the grammar.

Using an Overly General Grammar

One might consider the above assumptions too strong: What if the exact structure of the grammar is not known? In this variation, we instead assume the learner has an overly general grammar that includes rules not present in the true grammar (see Figure 1). For example, rather than having only the option of expanding a transitive verb phrase to a verb followed by a noun phrase, the learner's grammar also has the possibility of expanding such a phrase to a noun phrase followed by a verb phrase. This simulates learning a grammar with unknown structure while maintaining a tractable hypothesis space (in this case, not knowing the word order in the language, but knowing the relevant syntactic classes). The extra rules give the learner a larger hypothesis space to consider, and our previous hypothesis space is the subset of the new space in which the probability of each of our newly added expansions was zero. By using an overly general grammar, we introduce much more ambiguity as to the structure of any given sentence. Thus, one might expect different results than in the previous simulation, where the number of possible derivations for any given sentence was relatively small.

The procedure for calculating representativeness in the case of an overly general grammar was very similar to the previous case. We again are considering representativeness for sentences with unknown structure, and thus used importance sampling to calculate the integral. The proposal distribution for sampling was modified so that expansions with the added rules (not present in the true grammar) would be considered. We again used a Dirichlet distribution, but the parameters were equal to ten times the true parameters plus one. Thus, rules that had zero probability in the true grammar had a parameter of one in the Dirichlet prior.

As shown in Figure 4, even with a grammar with extra rules, the results are very similar to the previous simulation. Optimizing the corpora has the strongest effect when the complexity is somewhat limited, but for the greatest representativeness, it is still best to use a corpus with greater complexity. This result suggests that even if a learner does not know the true grammar, it is still better to provide a corpus of full complexity rather than a "starting small" corpus. However, several concerns remain. The way in which we generalized the grammar was limited to switching the orders of verb phrases and noun phrases. This adds significant ambiguity to the grammar, but is not equivalent to considering any arbitrary grammar. For example, one could imagine a grammar that had over-general rules for producing relative clauses. In that case, it is still unclear whether representativeness in a corpus of severely limited complexity could equal the representativeness of a more complex corpus. To fully explore the problem, we would need a tractable way to consider all (infinite) possible grammars that could produce the data.

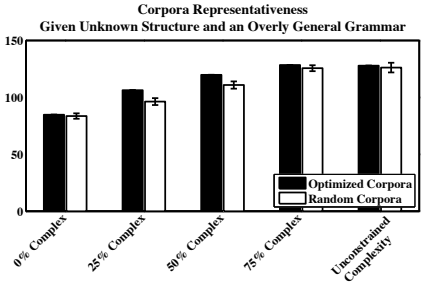

Figure 4: Representativeness of corpora with unknown structure using a grammar with extra rules.

Representativeness with N-Grams

The above simulations assume the learner learns a PCFG, but existing neural network models formulate language learning as learning to predict the next word based on previous words. This corresponds to a model where the learner learns distributions over n-grams rather than rules, and thus we can apply the same mathematical tools to analyze the representativeness of corpora according to an n-gram model.

To calculate representativeness, we can use exact counts as in the first simulation. An n-gram is a sequence of two (bigram) or three (trigram) words, and we assume a language model that estimates the probability of the next word given the previous one or two words. Our target distribution is now the correct proportions for each n-gram, which we estimate by computing the n-grams on the large corpus from which the other corpora were drawn.

Despite the fact that the model of the language has changed significantly, similar results hold in this case as in the other cases. Figure 5 shows the representativeness of the same corpora used in the other simulations with respect to n-grams: the random corpora of full complexity are still more representative than optimized corpora with limited complexity.

Summary

In our analysis, we have shown that starting small limits the degree of representativeness of the corpora and that the effect on representativeness is greatest when the limitations are particularly severe. These results hold regardless of the variations we considered. While our task is not exactly the same as in Elman (1993) or Rohde and Plaut (1999), it has bearing on this debate. From a computational level perspective, the only

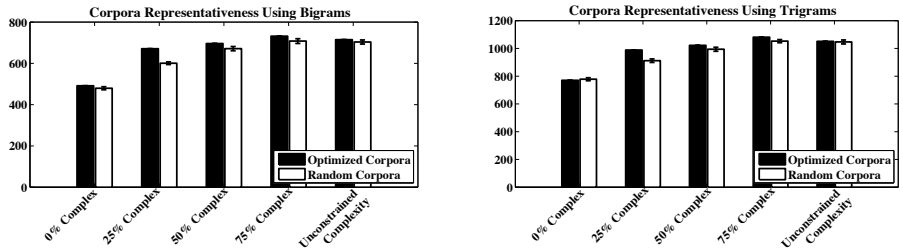

Figure 5: Representativeness of corpora using an n-gram model.

concern for such learning is whether the corpora are representative, and we have shown that starting small (at least in the extreme form) is not compatible with maximizing representativeness. However, if for mechanistic reasons one needs to start small, the above results suggest that starting "smaller" can result in similar representativeness in an optimized corpus to that of a random corpus of full complexity.

Discussion

We have shown how the concept of Bayesian representativeness can be applied to language in order to characterize an optimal sample and presented a case study of how representativeness changes with constraints on the sample. Mathematically, the Bayesian representativeness of language structures matches our intuitive sense of representativeness: a sample of language is most representative if the actual number of occurrences of each structure matches the expected number. While we cannot give a closed form expression for the representativeness of a corpus where the sentences structures are not given, simulations show that the trends concerning representativeness given constraints on complexity hold for these corpora as well. Finally, it is suggestive that given a grammar with overly general rules, we still find a disadvantage for corpora of limited complexity.

Our results suggest that if there is a beneficial effect of starting small, it is not located at the computational level: the statistical evidence a corpus provides in favor of the target language falls off as its complexity deviates from the complexity of the language. However, our results do show how it might be possible to start small in response to mechanistic information-processing constraints and still not impede learning, as it is possible to construct limited-complexity corpora that provide as much evidence as a random sample from the language. While suggestive, we note that these conclusions are tempered by the models we considered, and in particular the space of alternative hypotheses we allow the learner.

Overall, our analysis provides insight into what optimal linguistic input should look like in several interesting cases. A variety of next steps are possible. First, a more detailed exploration of the nature of an optimal sample given unknown rules would illuminate whether the preliminary results we have found hold given a larger space of possible grammars. Additionally, comparing our theoretical results to actual corpora of language acquisition would indicate whether child-directed speech is more representative than randomly selected adult-directed speech. This would suggest a pedagogical role for child-directed speech. This work provides a foundation for addressing these more advanced questions.

Acknowledgements. This work was supported by a Graduate Research Fellowship and grant number SES-0631518 from the National Science Foundation.

References

Baker, J. (1979). Trainable grammars for speech recognition. In J. J. Wolf & D. H. Klatt (Eds.), *Speech Communication Papers presented at the 97th Meeting of the Acoustical Society of America* (pp. 547–550). MIT, Cambridge, Massachusetts.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–224.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development: Some simple relationships. *Journal of Child Language*, *6*(3), 423–443.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: The less is more hypothesis. In J. M. Mead (Ed.), *The Proc. of the 11th West Coast Conference on Formal Linguistics.* Stanford, CA: CSLI.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development*, *73*(2), 418–433.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45*(3), 337–374.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). Dept. of Computer Science, University of Toronto.

Newport, E. L., Gleitman, L. R., & Gleitman, H. (1977). Mother I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. Snow & C. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 31–49). Cambridge, England: Cambridge University Press.

Pine, J. M. (1994). The language of primary caregivers. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition* (pp. 15–37). Cambridge, England: Cambridge University Press.

Rohde, D. L. (2003). *The simple language generator: Encoding complex languages with simple grammars* (Tech. Rep.). Department of Brain and Cognitive Science, MIT.

Rohde, D. L., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*, 67–109.

Sherrod, K. B., Friedman, S., Crawley, S., Drake, D., & Devieux, J. (1977). Maternal language to prelinguistic infants: Syntactic aspects. *Child Development*, *48*(4), 1662–1665.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, *43*(2), 549–565.

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 84–98).