Bulletin of Mathematical Biology (2004) **66**, 651–662
doi:10.1016/j.bulm.2003.09.007

Language Learning: How Much Evidence Does a Child Need in Order to Learn to Speak Grammatically?

KAREN M. PAGE

Department of Computer Science, Bioinformatics Unit,
University College London,
Gower Street,
WC1E 6BT London,
UK

E-mail: k.page@cs.ucl.ac.uk

In order to learn grammar from a finite amount of evidence, children must begin with in-built expectations of what is grammatical. They clearly are not born, however, with fully developed grammars. Thus early language development involves refinement of the grammar hypothesis until a target grammar is learnt. Here we address the question of how much evidence is required for this refinement process, by considering two standard learning algorithms and a third algorithm which is presumably as efficient as a child for some value of its memory capacity. We reformulate this algorithm in the context of Chomsky's 'principles and parameters' and show that it is possible to bound the amount of evidence required to almost certainly speak almost grammatically.

1. INTRODUCTION

A basic problem in the study of language is how children acquire grammar. In other words, how do they learn to distinguish sentences whose structures are allowable in their language from those which are not? Chomsky argued, and indeed it is logically provable (Gold, 1967), that if children start with absolutely no preconceptions about what is allowable, then they will not be able to infer the correct grammar from a finite amount of evidence (in the form of example grammatical sentences heard). Thus they must start with some built-in restrictions on the types of allowable structures. Chomsky dubbed these preconceptions 'Universal Grammar' (UG) (Chomsky, 1965). Children are not, however, born with a complete grammar hard-wired, as can be seen from their ability to learn a variety of different grammars, corresponding to different languages, as well as the mistakes they make in their early language usage. Thus in their early years they refine their hypothesis from that dictated by UG to a fully developed grammar. Here we study the amount of evidence required for this refinement process, as decisions are made in response to hearing positive evidence.

0092-8240/04/040651 + 12 $30.00/0

Three factors are of importance in estimating the amount of evidence required. These are (1) the initial assumptions that the child has (UG), (2) the learning algorithm that the child employs in the refinement process and (3) the distribution of the evidence the child hears.

We assume that a child's UG corresponds to a finite set of possible full grammars, between which the child must choose. [It is, in fact, also possible for the child to learn from an infinite set, provided that she/he has a sufficiently restrictive prior expectation of the probability distribution on the component grammars. We do not consider this possibility here (Vapnik, 1998).] Grammar learning then corresponds to whittling down the number of possible grammars in response to evidence heard, until the child is left with a single grammar.

In this context the important features of the distribution of evidence are the proportions of sentences heard (which are examples of the target/parental grammar) which are contained in each of the child's hypothesis grammars and their various intersections. Thus we need to define a measure on the set of sentences generated by the target grammar to determine the size of the intersection of this grammar with any other grammar/intersection of grammars.

Finally, considering the learning algorithm that the child employs, in this paper we consider two standard algorithms, the batch and memoryless learners (Niyogi, 1998), frequently used as examples because of their simplicity and because the batch learner is presumably about as fast a learner as could be employed by the child and the memoryless learner is about as unsophisticated as one could imagine. We also discuss a new algorithm which we show can be about as fast as the batch learner, if it makes extensive use of memory, and about as slow as the memoryless learner, if it does not. Thus this last algorithm is likely to learn at the rate of a child for some value of its memory capacity.

We estimate in each case how much evidence will be required for convergence to the target grammar. The amount of evidence needed depends crucially on the sizes of overlaps between grammars.

We reformulate the third algorithm in the context of 'principles and parameters' (Chomsky, 1981), as a parameter-setter. In this framework, the overlap between grammars depends on the frequency of sentences which determine each parameter. We derive results on the amount of evidence needed to learn the target grammar when certain assumptions are made about the frequency of usage of the parameters. For a general parameter frequency distribution, we show that it is possible to bound the amount of evidence needed to almost certainly produce a given proportion (almost one) of sentences grammatically correctly.

2. LEARNING ALGORITHMS

First, we describe the three learning mechanisms that we will study. We assume that the UG consists of a set of n grammars. The memoryless learner starts by

choosing one of these at random. On hearing a sample sentence from a teacher, the learner stays with the initial grammar if the sentence heard is consistent with that grammar, otherwise she/he picks a different grammar at random. This process is then repeated with the new grammar. Once the learner has picked the target grammar, she/he will never leave it, since all sentences should be consistent with this grammar (we assume, for now, that the teacher speaks perfectly grammatically). If the learner has picked another grammar then eventually she/he should hear evidence which conflicts with this grammar and move on. Thus the target grammar is the only fixed point of the process. Since random picking of the new grammar implies that there is always a nonzero probability of picking the target grammar, the learner will eventually converge on the target. The learner has no memory and so at any stage can pick a grammar which she/he has previously rejected.

The batch learner remembers all the evidence presented to him/her and rejects hypothesis grammars if she/he has ever heard evidence which conflicts with the grammar. Thus in time the set of hypothesis grammars decreases until the learner has heard evidence which conflicts with all but the target grammar and so has learnt this grammar. In order to assign a grammar to the learner at any given time, we assume that she/he picks one at random from the remaining set.

We call the last type of learner the 'pairwise learner'. We list the hypothesis grammars $G_1, \ldots, G_n$. On receiving evidence, she/he compares G_1 with G_2, G_3 with $G_4, \ldots, G_{n-1}$ with G_n until at least one of each pair is rejected. Then, relabelling in the set of remaining grammars (which has size $\leq n/2$), she/he repeats the process over and over, each time at least halving the number of remaining grammars, until only the target is left. Once again, at any one time, the learner's grammar is chosen at random from the remaining hypothesis grammars.

3. The Amount of Evidence Required for a Given Learner to Converge on the Target Grammar

3.1. ***Memoryless learner.*** First we consider the memoryless learner. The probability that the learner has not yet converged to the target grammar is given by

$$\begin{aligned}
&\sum_{k=1}^{M+1} P(\text{have rejected } k-1 \text{ grammars}) \\
&\qquad \times P(\text{target grammar is not one of the } k \text{ grammars picked first}) \\
&\qquad = \sum_{k=1}^{M+1} \left(\frac{n-1}{n}\right)^k \times P(\text{have rejected } k-1 \text{ grammars}), \qquad (1)
\end{aligned}$$

where M is the number of sentences that the child has heard.

Now the worst case scenario will be if all the overlaps are large. Let γ_i be the overlap between G_i and the target grammar and let $\mu = 1 - \gamma_{\max}$, where $\gamma_{\max} = \mathrm{Max}_i \gamma_i$. Then $\gamma_i \geq 1 - \mu, \forall i$. So considering the worst case, when $\gamma_i = 1 - \mu, \forall i$, the probability that the learner has rejected r grammars after hearing M sentences is given by

$$P(\text{rejected } r \text{ after } M) = \binom{M}{r} \mu^r (1-\mu)^{M-r}. \tag{2}$$

Thus

$$P(\text{not reached target}) \tag{3}$$

$$= \sum_{k=1}^{M+1} \left(\frac{n-1}{n}\right)^k \binom{M}{k-1} \mu^{k-1} (1-\mu)^{M-k+1} \tag{4}$$

$$= \frac{n-1}{n} \sum_{k=0}^{M} \binom{M}{k} \mu^k \left[\frac{n-1}{n}\right]^k (1-\mu)^{M-k} \tag{5}$$

$$= \frac{n-1}{n} \left[(1-\mu) + \frac{n-1}{n}\mu\right]^M \tag{6}$$

$$= \frac{n-1}{n} \left[1 - \frac{1}{n}\mu\right]^M. \tag{7}$$

Now we require that this probability should be less than δ; this will be true if

$$\left[1 - \frac{\mu}{n}\right]^M < \frac{n}{n-1}\delta. \tag{8}$$

Therefore, in order for the probability of not having reached the target to be less than δ, we require approximately that

$$M > \frac{\log \delta}{\log\left[1 - \frac{\mu}{n}\right]} \approx \frac{\log \delta}{-\frac{\mu}{n}} = n\frac{\log \frac{1}{\delta}}{\mu}. \tag{9}$$

Thus if all the overlaps are bounded away from 1, the memoryless learner will learn the target after $O(n)$ sentences. Otherwise, define $\alpha(n)$, by $\alpha(n) = 1/(1 - \gamma_{\max}(n))$ [for a uniform distribution of overlaps $\alpha(n) \sim n$ (Komarova and Rivin, 2001)]. Intuitively, $\alpha(n)$ is the average number of sentences that one has to sample in order to find one which distinguishes between the two most closely related grammars. The memoryless learner can learn the target grammar after hearing $O(n\alpha(n))$ sentences.

By considering the best case scenario, $\gamma_i = 0, \forall i$, we can see that we require $M > n \log(1/\delta)$, so the memoryless learner is always liable to require at least $O(n)$ sentences to learn the target.

Now

$$E(M) = \sum_{r=1}^{\infty} P(\text{target is } r\text{th grammar picked}) \times E(\text{number of sentences required to reject first } r-1 \text{ grammars} \mid \text{they are not target}) \tag{10}$$

and

$$E(\text{number of sentences required to reject } G_i) = \sum_{k=1}^{\infty} k(1-\gamma_i)\gamma_i^{k-1} = \frac{1}{1-\gamma_i}. \tag{11}$$

Hence

$$E(\text{number of sentences required to reject first grammar} \mid \text{it is not the target}) = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{1-\gamma_i} \tag{12}$$

where, without loss of generality, we take G_n to be the target grammar. The same is true for the rejection of the second, third, ..., rth grammars picked. Thus

$$E(M) = \sum_{r=1}^{\infty}\frac{1}{n}\left(\frac{n-1}{n}\right)^{r-1}(r-1)\left[\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{1-\gamma_i}\right] \tag{13}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n-1}\frac{1}{1-\gamma_i}\sum_{r=0}^{\infty} r\left(\frac{n-1}{n}\right)^{r-1} \tag{14}$$

$$= \frac{1}{n^2}\frac{1}{\left(1-\frac{n-1}{n}\right)^2}\sum_{i=1}^{n-1}\frac{1}{1-\gamma_i} \tag{15}$$

$$= \sum_{i=1}^{n-1}\frac{1}{1-\gamma_i}. \tag{16}$$

This is clearly less than or equal to $O(n\alpha(n))$, see also Rivin (2002). If we know more about the distribution of overlaps, then we may be able to obtain a more definite bound.

3.2. ***Batch learner.*** Now we consider the batch learner. The probability that the learner has not rejected all but the target grammar after hearing M sentences is equal to the probability that at least one of the incorrect grammars is still remaining. Now the probability that grammar i is remaining is given by γ_i^M so the probability

that there exists an incorrect grammar remaining is given by

$$1 - \prod_{i=1}^{n-1} [1 - \gamma_i{}^M]. \tag{17}$$

Thus if we want the probability of having learnt the target grammar to exceed $1-\delta$ then we must demand that

$$1 - \prod_{i=1}^{n-1} [1 - \gamma_i{}^M] < \delta \tag{18}$$

$$\Leftrightarrow \prod_{i=1}^{n-1} [1 - \gamma_i{}^M] > 1 - \delta \tag{19}$$

$$\Leftrightarrow \prod_{i=1}^{n-1} \log[1 - \gamma_i{}^M] > \log(1 - \delta) \approx -\delta. \tag{20}$$

Now $\log[1 - \gamma_i{}^M] > \log[1 - \gamma_{\max}^M]$, where $\gamma_{\max}(n)$ is the maximal proportion overlap between an incorrect hypothesis grammar and the target grammar, which we assume to be almost 1; so the above inequalities will hold if

$$(n-1)\log[1 - \gamma_{\max}^M] > -\delta. \tag{21}$$

This will only be true if $\gamma_{\max}^M$ is small and hence if

$$\gamma_{\max}^M < \frac{\delta}{n-1} \tag{22}$$

$$\Leftrightarrow M > \frac{\log(\delta) - \log(n-1)}{\log(\gamma_{\max})}. \tag{23}$$

Thus the batch learner requires $O(\log n/|\log(\gamma_{\max})|)$. Now $\gamma_{\max} \in (0, 1)$, so $|\log[\gamma_{\max}]| > 1-\gamma_{\max}$. Thus a batch learner requires at most $O(\log n/(1-\gamma_{\max})) = O(\alpha(n)\log n)$ sentences to learn the grammar.

If we know more about the distribution of overlaps, we can obtain a tighter bound. Let us assume that the overlaps are independently uniformly distributed in [0, 1].

Now from (18), we have

$$\sum_{i=1}^{n-1} \log[1 - \gamma_i{}^M] > -\delta. \tag{24}$$

This implies that all the $\gamma_i{}^M$ must be very small and hence

$$\log[1 - \gamma_i{}^M] \approx -\gamma_i{}^M. \tag{25}$$

So,

$$\sum_{i=1}^{n-1} {\gamma_i}^M < \delta. \tag{26}$$

Now

$$E\left(\sum_{i=1}^{n-1} {\gamma_i}^M\right) = (n-1)E({\gamma_i}^M) = (n-1)\int_0^1 x^M dx = \frac{(n-1)}{M}, \tag{27}$$

so

$$M > \frac{n-1}{\delta}. \tag{28}$$

Thus with a uniform distribution of overlaps, the learner requires $O(n)$ sentences to learn the target grammar.

3.3. ***Pairwise learner.*** Lastly, we consider the pairwise learner. First it is clear that if the learner has an unlimited memory capacity and can reuse sentences used to compare early pairs in the later comparisons, then the learner is equivalent to a batch learner, since all comparisons between grammars are effectively made at once.

Secondly, we consider what happens when the learner has no memory for sentences and can only use a given sentence to compare the two grammars that she/he is currently considering. In this case, she/he must take $n-1$ comparisons between grammars. The comparison between G_i and G_j will be over with probability $1-\delta$ after r sentences where

$$\gamma_i^r * \gamma_j^r < \delta \tag{29}$$

$$r > \frac{\log(\delta)}{(\log(\gamma_i) + \log(\gamma_j))} \tag{30}$$

this will be the case if $r > \frac{\log(\delta)}{\log(\gamma_{\max})}$. Thus the target grammar will be learnt in $O(n/\log(\gamma_{\max}))$ sentences. As we have been before this is (at most, with equality if α is large) $O(n\alpha(n))$.

The expected number of sentences required to make the comparison between G_i and G_j is clearly less than or equal to the expected number of sentences to reject G_i, where (without loss of generality) G_i has an overlap with the target which is less than or equal to that of G_j. Now the expected number of sentences to reject G_i is $\frac{1}{1-\gamma_i}$ and hence the expected number of sentences required to perform the $n-1$ comparisons necessary for the pairwise algorithm

$$E(M) \leq \sum_{i=1}^{n-1} \frac{1}{1-\gamma_i}, \tag{31}$$

since if G_i appears more than once as the minimally overlapping grammar of a pair

then it means that it has been compared more than once with a more overlapping grammar, which means that it has not been rejected at least once in such a comparison and hence it substitutes for another grammar in the sum which has a larger overlap and hence would have made a larger contribution to the sum.

Thus the expected number of sentences required for the pairwise algorithm with no memory for sentences to converge to the target grammar is less than or equal to the number required by the memoryless algorithm.

Now the expected number of sentences required to compare grammars i and j is given by $\frac{1}{1-\gamma_i\gamma_j}$ and hence

$$E(M) = \sum_{\text{pairs } (i,j) \text{ compared}} \frac{1}{1-\gamma_i\gamma_j}. \tag{32}$$

Suppose the overlaps are close to one and set $\gamma_i = 1-\epsilon_i$ and $\gamma_j = 1-\epsilon_j$, then $\frac{1}{1-\gamma_i\gamma_j} \approx \frac{1}{1-(1-\epsilon_i-\epsilon_j)} = \frac{1}{\epsilon_i+\epsilon_j}$. Now if γ_i is close to 1, then the probability that it will be rejected in a comparison with a grammar with a significantly smaller overlap with the target is infinitesimal. Hence, with probability almost 1, the comparison which results in the rejection of G_i will require $O(\frac{1}{\epsilon_i})$ sentences. Thus $E(M)$ will be of the order of

$$\sum_{i=1}^{n-1} \frac{1}{\epsilon_i} I[\gamma_i \approx 1] + \sum_{i=1}^{n-1} 1 I[\gamma_i \text{ bounded away from 1}] = O\left(\sum_{i-1}^{n-1} \frac{1}{1-\gamma_i}\right), \tag{33}$$

where $I[\ldots]$ is the indicator function (taking value 1 if $\ldots$ is true and 0 otherwise). Hence the pairwise learner (with no memory for sentences) takes the same order of number of sentences to converge to the target as the memoryless learner does.

It has thus been shown that the pairwise algorithm can be as fast as the batch or as slow as the memoryless learner depending on its memory capacity and so for some capacity it is presumably as fast as a real human learner.

4. Parameter Setting if the Parameter Frequencies Follow Certain Specific Distributions

We assume that the grammar is determined by m binary parameters, $a_1, \ldots, a_m$. Thus there are $n = 2^m$ candidate grammars. We assume that a given sample sentence determines a_i with probability α_i, where $1 \geq \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n > 0$. Thus the parameters are listed in descending order of frequency. We assume that if the setting of parameter a_j depends on the prior setting of a_i, then a_j will be less frequent than a_i and hence $j > i$. Thus more accurately, we define α_i to be the probability that a random sentence determines a_i given that the values of $a_1, \ldots, a_{i-1}$ are known. We thus set parameters in turn starting with a_1. [This fits in with what is

observed in real children's learning of grammar. 'It would make sense if children, too, instinctively work their way down the (parameter) hierarchy, ... overall these results are encouraging for the view that the parameter hierarchy provides a logical flowchart that children use in the process of language acquisition' (Baker, 2001).] Each parameter-setting corresponds to a halving of the number of remaining grammars and thus the algorithm is equivalent to the pairwise algorithm where the pairs in the first round of comparisons correspond to grammars differing only in the first parameter, in the second round to grammars differing in the second parameter (and maybe the first, but not the others) and so on.

Setting the first parameter takes a mean number of sentences, $E(s_1)$, given by

$$E(s_1) = \sum_{k=1}^{\infty} k P(\text{first } k-1 \text{ sentences do not determine } a_1, \text{ but } k\text{th does})$$

$$= \sum_{k=1}^{\infty} k(1-\alpha_1)^{k-1}\alpha_1 = \frac{1}{\alpha_1}. \tag{34}$$

Similarly, setting parameter a_i takes a mean number of sentences given by $E(s_i) = \frac{1}{\alpha_i}$. Thus the expected number of sentences required to learn the target grammar is given by

$$E(M) = \sum_{i=1}^{m} \frac{1}{\alpha_i}, \tag{35}$$

for any distribution of frequencies. In particular if the frequency of sentences determining a given parameter, given that earlier parameters have already been learnt, is given by a constant ν, then $E(M)$ is given by $m/\nu = O(\log n)$. If the parameter frequencies follow Zipf's law, that is the $\alpha_i = \kappa/i$, then $E(M) = \frac{1}{\kappa}\sum_{i=1}^{m} i = \frac{1}{\kappa}\frac{m}{2}(m-1) = O(m^2) = O((\log n)^2)$.

5. The Amount of Evidence Required for a Child to Learn to Speak Almost Grammatically for General Parameter Frequency Distributions

Suppose that a learner has heard M sentences from his/her teacher. We assume now for simplicity that the learner has batch-style unlimited memory capacity, so that although she/he must fix parameters $a_1, \ldots, a_{r-1}$ before parameter a_r, she/he can reuse the same sentences as were used to fix $a_1, \ldots, a_{r-1}$ in order to fix a_r. She/he will have learnt the first $0 \leq m' < m$ parameters with probability

$$1 - P(\text{have not received a sentence determining } a_i \text{ for some } i \leq m')$$

$$\geq 1 - \sum_{i=1}^{m'} (1-\alpha_i)^M \geq 1 - m(1-\alpha_{m'})^M. \tag{36}$$

So a learner will have learnt the first m' parameters with probability $> 1 - \delta$ if $m(1 - \alpha_{m'})^M < \delta$.

The probability that a learner who has learnt the first m' parameters (and only these) speaks a randomly selected sentence correctly is

$$\geq (1 - \alpha_{m'+1}) \prod_{i=m'+2}^{m} (1 - \beta_i), \tag{37}$$

where β_i is the probability that a random sentence determines a_i given that $a_1, \ldots, a_{m'}$ are known. Now all sentences which determine a_i given that $a_1, \ldots, a_{m'}$ are known also determine a_i given that $a_1, \ldots, a_{i-1}$ are known, provided that $i - 1 \geq m'$. Hence, $\alpha_i \geq \beta_i$ for $i \geq m' + 1$ and so the probability that a learner who has learnt the first m' parameters (and only these) speaks a randomly selected sentence correctly is

$$\geq \prod_{i=m'+1}^{m} (1 - \alpha_i) \geq (1 - \alpha_{m'+1})^m. \tag{38}$$

Thus a learner who has learnt the first m' parameters will speak at least a proportion $1 - \epsilon$ correctly if

$$(1 - \alpha_{m'+1})^m > 1 - \epsilon. \tag{39}$$

This will be true if

$$\log(1 - \alpha_{m'+1}) > 1/m \log(1 - \epsilon) \approx -\epsilon/m, \tag{40}$$

that is if $\alpha_{m'+1} < \epsilon/m$. Thus a learner will produce at least a proportion $1 - \epsilon$ of sentences correctly if $\forall r$ such that $\alpha_r \geq \epsilon/m$, the learner has learnt parameter a_r. This will be true with probability $1-\delta$ if $m(1-\epsilon/m)^M < \delta$. Thus with probability $> 1-\delta$ a learner will make mistakes in a proportion $< \epsilon$ of sentences provided that

$$M > \log\left(\frac{\partial}{m}\right) \Big/ \log\left(1 - \frac{\epsilon}{m}\right) \approx \frac{m}{\epsilon}[\log m - \log(\delta)]. \tag{41}$$

Thus a learner will probably produce almost all sentences correctly after hearing $O(\log n \log(\log n))$ sentences.

6. CONCLUSIONS

We have looked at the amount of evidence that a child requires to learn a target grammar from a (finite) set of hypothesis grammars (UG). We consider three learning algorithms: the memoryless learner, the batch learner and the 'pairwise learner'. The memoryless learner is about as unsophisticated an algorithm as we can imagine and requires $O(\alpha(n)n)$ sentences to choose from n hypothesis grammars, where $\alpha(n) = 1/(1-\gamma_{\max}(n))$ and $\gamma_{\max}(n)$ is the maximal proportion overlap between the grammars. The batch learner, which is as fast a learner as could be employed, requires $O(\alpha(n) \log n)$ sentences. The pairwise algorithm can be about

as slow as the memoryless learner or as fast as the batch learner, depending on the memory capacity. Thus for some value of the memory capacity, we assume that it will learn at the same speed as a child does.

We consider the pairwise algorithm in the context of principles and parameters. It is equivalent to a parameter-setter. We find that if the frequency with which parameter i is determined by a random sentences tends to zero as i becomes large, then the learner will tend with probability $1 - \delta$ to produce grammatically correct sentences with probability $1 - \epsilon$ when $\approx m/\epsilon[\log m - \log \delta]$ sentences have been heard. Here m is the number of parameters and hence is $\log_2 n$, where n is the number of hypothesis grammars. Note that this is an example of 'Probably almost correct' learning (Valiant, 1984; Vapnik, 1998). If the frequency with which parameter i is determined by a random sentence is given by a constant ν, then the learner will have converged on the target grammar after hearing $O(\log n)$ sentences. If instead the frequencies follow Zipf's law, then the learner will have converged on the target grammar after hearing $O((\log n)^2)$ sentences.

It would be interesting to look at the distribution of frequencies of parameters from real language data and see what conclusions could be drawn about the amount of evidence need to learn these real languages.

It would also be interesting to investigate the effects on not-quite-perfect grammar learning on the population dynamics of the grammar. [For a review of the population dynamics of grammar see Nowak *et al*. (2002).] For instance, a heterogeneity in the values of rare parameters could presumably lead to slow language change over time. Also, if different members of the population have subtly different grammars, then it clearly matters who your teachers are. It would be interesting to study the dynamics of learning from multiple teachers. This would necessitate a device for the learner to reconcile conflicting evidence. Such evidence could lead in the cases of batch or pairwise learners to an empty set of remaining grammars, thus halting the learning process. It would not, however, halt the implementation of a memoryless algorithm, but would mean that the learner would never converge upon a grammar.

Here we consider only learners that choose between a restricted set of hypotheses. Other learning algorithms akin to neural networks have been applied to language learning. It is claimed that such algorithms, with much less structured UGs, may be capable of learning language if the evidence presented to them is of a very simplistic nature. It is possible that the kind of evidence presented to children in very early language acquisition (so-called 'Motherese') is of this nature. Thus the question of how much grammar must be already present in the infant's brain remains open.

ACKNOWLEDGEMENTS

KMP would like to acknowledge support from the Joint Research Councils (EPSRC, BBSRC and MRC, grant number GR/R47455). She would also like to thank Natasha Komarova, Martin Nowak and Igor Rivin for helpful comments.

REFERENCES

Baker, M. C. (2001). *Atoms of Language*, NY: Basic Books.
Chomsky, N. (1965). *Aspects of the Theory of Syntax*, Cambridge, MA: MIT Press.
Chomsky, N. (1981). *Lectures on Government and Binding*, Dordrecht: Reidel.
Gold, E. M. (1967). Language identification in the limit. *Inf. Control* **10**, 447–474.
Komarova, N. L. and I. Rivin (2001). Mathematics of learning. Preprint math.PR/0105235 (at http://lanl.arXiv.org).
Niyogi, P. (1998). *The Informational Complexity of Learning: Perspectives on Neural Networks and Generative Grammar*, Boston: Kluwer Academic Publishers.
Nowak, M. A., N. L. Komarova and P. Niyogi (2002). Computational and evolutionary aspects of language. *Nature* **417**, 611–617.
Rivin, I. (2002). The performance of the batch learning algorithm. Preprint cs.LG/0201009 (at http://lanl.arXiv.org).
Valiant, L. G. (1984). A theory of the learnable. *Commun. ACM* **27**, 436–445.
Vapnik, V. N. (1998). *Statistical Learning Theory*, NY: Wiley.

Received 30 July 2003 and accepted 26 September 2003