J. Child Lang. **31** (2004), 954–958. Printed in the United Kingdom
© 2004 Cambridge University Press

DISCUSSION

DOI: 10.1017/S0305000904006440

Learnability, stochastic input, and connectionist networks: a response to Brian MacWhinney's 'A multiple process solution to the logical problem of language acquisition'

DOUGLAS L. T. ROHDE
Massachusetts Institute of Technology
dr@tedlab.mit.edu

I would like to begin by clarifying some technical points regarding learnability and Gold's theorem and will work my way around to advocating the power of distributed connectionist networks. I may wind up repeating much of what MacWhinney has already said, but I hope to do so in a manner that clarifies the status of and relationship between his seven solutions to the language acquisition problem.

MacWhinney seems to conflate finite languages, which are learnable under Gold's formulation, with regular languages, which can be recognized or produced by finite state automata. To be clear, a finite language simply consists of a finite set of sentences. Such a language is easy to learn in Gold's model because the learner can simply assume that all sentences it has observed are in the language and all others are not (a primitive conservatism). Eventually, the learner will have observed the full set and will thus have learned the language.

Finite state machines, on the other hand, produce regular languages, which are potentially infinite and which, as a class, are a superset of the finite languages. For example, the infinite language of all sentences with an even number of words is regular, as is the language consisting of every possible sentence. However, the regular or 'finite-state' grammars are not, in fact, learnable from positive evidence under Gold's model of learning. Even given a set of positive and negative examples, it is NP-hard, and thus presumably computationally intractable, to infer the smallest finite automaton consistent with those examples.[1] No language class that includes all of the finite languages and at least one infinite language is learnable in Gold's sense. Therefore, the learnability of natural language is not improved by expressing

[1] An NP-hard problem has no known solution that runs in time that is less than exponential in the size of the input. This makes it computationally intractable for all but the smallest problems.

that language as a left-associative grammar (Hausser, 1999) or a finite-state grammar with limited continuations (Reich, 1969).

Learnability has been shown under Gold's model for some restricted classes of grammars that, unlike the regular and context-free languages, do not include all possible finite languages. For example, parenthesis languages (McNaughton, 1967), k-reversible regular languages (Angluin, 1982), and the class of k-valued languages (Kanazawa, 1998) are learnable under Gold's model. However, some of these language classes cannot adequately characterize the natural languages. For example, the fact that natural language contains verbs that share some, but not all, argument structures, and the fact (if we choose to believe it) that noun phrases or other verb arguments can be arbitrarily long, violates the assumptions of k-reversibility.

On the other hand, Shinohara (1994) has shown that the class of context-sensitive languages defined by at most *n* productions is learnable. These are sufficient for characterizing natural language if one is willing to place an *a priori* bound on the complexity of a natural grammar, which seems entirely reasonable. However, the inference methods invoked in this and similar learnability proofs are, by and large, computationally infeasible, requiring the learner to enumerate all possible languages. Likewise, inferring k-valued languages is NP-hard, even with structured examples (Florêncio, 2000). The language classes learnable under Gold's model are either overly constrained or are not, as far as we know, learnable in a computationally practical manner.

Gold's learning model is difficult largely because of the very loose constraints placed on the information source, which is required only to produce every sentence eventually. But the source could withhold entire structures from the language for an indefinite, albeit not infinite, time. As a result, there is no useful guarantee that would justify the eventual decision that if the learner hasn't heard something yet, that thing is probably ungrammatical. Mechanisms such as competition and blocking are inapplicable under Gold's model because they rely on just such a guarantee.

Conservatism is a property of virtually all successful learning strategies in Gold's framework and the language classes that aren't learnable are precisely those for which conservatism is either not possible or doesn't help. So conservatism in itself is not a solution to the learnability problem, and, as MacWhinney points out, we know that children are not strictly conservative. Conservatism must at times be abandoned in lieu of generalization, so I would remove conservatism from the list of solutions and demote it to a rule of thumb. In a Bayesian framework, conservatism is that aspect of the system that maximizes fit with the data, while generalization strives to maximize model simplicity.

Truly solving the logical problem of language acquisition requires altering the assumptions of the learning framework. MacWhinney hit upon one aspect of this in his suggestion that the end-state criterion should not require settling on a perfect grammar, but a close approximation to one (Horning, 1969). However, the other critical part of this reformulation is the assumption that language is stochastic and that the frequency of sentences or structures observed by a learner are governed by a probability distribution. As a result, the past is a more or less accurate predictor of the future and the learner can reasonably assume that any structure not observed for a certain period of time is either ungrammatical or is sufficiently rare as to be insignificant.

In Rohde & Plaut (1999), we distinguished between INDIRECT negative evidence, which includes information gained from responses to a child's productions, such as the parents' misunderstandings or rephrasings, and IMPLICIT negative evidence, which exists only in the statistics of passively observed language. The usefulness of indirect negative evidence remains a matter of debate, but what we would call implicit negative evidence, MacWhinney's seventh solution, relies on the assumption of stochastic presentation, which is a much stronger constraint than that placed on Gold's text source. More rapidly applicable forms of competition or blocking rely on an assumption of a one-to-one mapping from meanings to words or structures. This, of course, assumes the presence of semantics, a critical half of the puzzle that is almost totally neglected in formal learning theory.

The introduction of probabilistic information raises some serious questions. Most notably, which statistics are to be monitored and how is this information used? Tracking the frequency of every possible complex contingency would be prohibitively expensive in terms of the memory and time required. But how can the learner find out if a contingency is important unless it is tracked? Linguists have long agreed that statistics gathered over surface forms, such as n-grams, cannot go very far in modelling natural language. Effective use of statistics requires generalization. To begin with, words must be clustered into classes, on the basis of surface co-occurrence statistics and meaning-to-utterance correspondences. Some statistics can then be recorded over classes, rather than over individual words, and further structure can be induced from the statistical relationships between classes in the input. What MacWhinney calls cue construction is one aspect of a very central process of structure-building that subserves and is driven by observed statistical information.

All of this seems quite difficult to realize in a discrete, symbolic system without extensive prior knowledge of the types of structures and relationships to expect. However, distributed connectionist networks may represent a solution. Such networks have demonstrated the ability to learn simple stochastic languages from positive presentation and, in doing so, to acquire representations of basic lexical categories that subserve additional learning (Elman, 1991).

A key distinction between distributed networks and localist networks or symbolic systems is that the internal representations learned by distributed networks inhabit a high-dimensional, continuous space. Through learning, representations with particular statistical similarities can be gradually pulled together along some or all of these dimensions to form functional clusters or categories. Subsequent learning involving one or more of these items will naturally generalize to similar items. However, such representations, unless overtrained, typically retain some of their uniqueness. If more subtle contingencies become apparent in the future, differently affecting the members of a category, the representations can be pulled apart again, possibly along orthogonal dimensions. Such representations can, under various pressures, effectively underly either graded or rule-like behaviour, both of which are required to model natural language syntax and semantics. We are just beginning to understand how the ability to induce structure from stochastic input might scale up to explain natural language learning without a reliance on extensive, detailed innate knowledge. A step in this direction is the recurrent, distributed network model presented in Rohde (2002), which is capable of learning, within a reasonable error tolerance, to comprehend and produce a fairly complex subset of English. This network primarily learns production on the basis of formulating predictions during comprehension, but can also use its comprehension system to refine productions by means of monitoring.

In summary, formal learning under Gold's model is even more difficult than MacWhinney suggests. However, adopting the alternative assumption that language is stochastic and need only be learned to a close approximation opens up a wealth of possible solutions to the learning problem. Distributed neural networks seem particularly well-adapted to learning from implicit negative evidence in a stochastic environment and are able to employ monitoring and exhibit such behaviours as competition and cue construction.

REFERENCES

Angluin, D. (1982). Inference of reversible languages. *Journal of the Association for Computing Machinery* **29**, 741–65.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* **7**, 195–225.

Florêncio, C. C. (2000). On the complexity of consistent identification of some classes of structure languages. In *Grammatical inference: algorithms and applications, 5th International Colloquium, ICGI 2000, Lisbon, Portugal, September 11–13, 2000; Proceedings* (**1891** 89–102). Springer, Berlin.

Hausser, R. (1999). *Foundations of computational linguistics: man-machine communication in natural language.* Berlin: Springer.

Horning, J. (1969). *A study of grammatical inference.* Stanford University, Computer Science Department.

Kanazawa, M. (1998). Learnable classes of categorial grammars. Unpublished doctoral dissertation, Stanford, CA.

McNaughton, R. (1967). Parenthesis grammars. *Journal of the Association for Computing Machinery* **14**, 490–500.

Reich, P. A. (1969). The finiteness of natural language. *Language* **45**, 831–43.

Rohde, D. L. T. (2002). A connectionist model of sentence comprehension and production. Unpublished doctoral dissertation, Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA.

Rohde, D. L. T. & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition* **72**(1), 67–109.

Shinohara, T. (1994). Rich classes inferable from positive data: length-bounded elementary formal systems. *Information and Computation* **108**, 175–86.