J. Child Lang. **31** (2004), 915–918. Printed in the United Kingdom

DISCUSSION

DOI: 10.1017/S0305000904006348

What can be learned from positive data? Insights from an 'ideal learner'. Commentary on 'A Multiple process solution to the logical problem of language acquisition' by Brian MacWhinney

NICK CHATER

Institute for Applied Cognitive Science, Department of Psychology, University of Warwick
nick.chater@warwick.ac.uk

MacWhinney's stimulating discussion suggests that there are many lines of argument that may address concerns raised by theorists who are concerned that there is a logical problem of language acquisition. This commentary argues: (1) that if the 'logical problem' applied to language, it would apply, with curious consequences to any learning by experience; (2) that the logical problem does not apply – given sufficient positive data from any reasonable language, language can be learned, in a probabilistic sense, by an 'ideal learner' using a simplicity principle; and (3) that a simplicity, or minimum description length, principle may provide a useful methodology for assessing claims concerning learnability of particular linguistic structures.

At a general level, one aspect of the so-called logical problem of language acquisition is that positive data alone may appear to be insufficient to allow a language to be learned. As MacWhinney notes, some theorists have interpreted Gold's (1967) results on learnability in the limit as indicating that language cannot be learned from positive data alone. Essentially, the problem is that, given positive data alone, it appears that the learner cannot recover from positing an overgeneral grammar – because the mere non-occurrence of a sentence cannot be evidence that the sentence does not occur. This is because given that language is infinite, and any corpus is finite, the overwhelming majority of legitimate sentences of the language do not occur.

This might suggest that the learner should be conservative – i.e. to postulate the smallest language consistent with the available data. But then we run into the opposite problem – that, unless subject to innate constraints to the contrary, the learner will simply choose the sentences of the finite corpus itself, as the shortest corpus.

MacWhinney notes various points of controversy concerning this type of argument. One issue is whether the child has access to, and can use, negative

data. A second issue is the precise application of Gold's results. A third issue is that there are positive results about learning from positive evidence that provide a counterweight to Gold's results. Specifically, learning in a probabilistic sense (rather than learning in the limit) is achievable in some specific types of language, given assumptions about how sentences are sampled from the language (e.g. Horning, 1969). We further explore this latter issue here.

First, note that, if learning from positive data is logically problematic, then this problem arises not merely for learning language, but for any aspect of learning from experience. For example, it applies to scientific enquiry – scientists have to build theories entirely on the basis of positive data – i.e. data that they obtain from their observations and experiments. They do not have access to 'negative' data, about what could not have occurred in those observations and experiments (they may, of course, have access to background knowledge drawn from other scientific domains; but the derivation of any such knowledge itself faces the same issue). Thus, any argument for nativism concerning language acquisition that is driven by the 'logical problem' will be paralleled by nativism concerning, say, cosmology or genetics. Thus it appears that, at best, the constraints that the 'logical problem' sets must be quite weak – or we may be faced with a Cosmology Acquisition Device to explain recent scientific advances, alongside the Language Acquisition Device postulated to explain children's learning.

Second, note that formal results have shown that, in a very general setting, positive evidence is sufficient to support learning (e.g. Horning, 1969; Solomonoff, 1978). Specifically, suppose that a stream of data is produced by any combination of computable and random factors (according to the assumption that cognition is computation, this will include any corpus of natural language). Then, an ideal learner can learn to predict each continuation of this corpus, with a FINITE sum-squared error between the predicted continuation and the true probabilities of each continuation over the entire infinite corpus (Solomonoff, 1978; see Li & Vitányi, 1997). This implies that the learner's expected level of error asymptotes to arbitrarily close to zero, given sufficient positive data. The computability restrictions aside, this result applies to any language, with any method sampling.

The ideal learner achieves this by applying a simplicity principle, such as is widely used in the study of perceptual organization (see, e.g. Chater, 1996; van der Helm, 2000). The strategy is to find the shortest description of the corpus so far; and to predict on the basis of that description. Whether the child actually uses such a principle is an interesting and open empirical question – but the existence of results concerning the ideal learner are enough to block a purely logical problem of language acquisition. Moreover, research I am currently conducting with Paul Vitányi indicates the same mathematical framework can be used to generate results concerning the learnability

of grammaticality judgements, language production, and even aspects of semantics.

The discussion so far has focused on the so-called logical problem of language acquisition at a general level. But this does not preclude the possibility that there are specific linguistic phenomena that children acquire without a sufficient evidential base, suggesting that this acquisition is constrained by innate, and perhaps language-specific, constraints.

Such arguments can only be dealt with in a piecemeal fashion; and our knowledge of what the child learns, given what input, will typically be too weak to provide them with a definitive resolution. MacWhinney provides an elegant summary of a range of sources of information and learning mechanisms that provide a counterweight to assumptions that specific phenomena are unlearnable in principle. This raises the question of whether there is some method of, however crudely, providing a quantitative analysis of the learnability of specific linguistic phenomena. One possibility, based on the simplicity principle used for our ideal learner above, is to employ the 'minimum description length' (MDL) principle: that the learner should prefer whichever linguistic structures provide the shortest description of the data (e.g. Barron, Rissanen & Yu, 1998). MDL, and a range of related mathematical and statistical ideas, provides a rigorous approach to inferring structure from data; and can be used to learn linguistic structure for corpora of language. The MDL principle is able to rule out overgeneral hypotheses, because these allow too large a class of possible sentences – and hence it is excessively costly (in terms of description length) to choose any particular sentence. The approach has been used to analyse language learnability and acquisition in a number of contexts (e.g. Ellison, 1992; Brent & Cartwright, 1996; Grünwald, 1996; Clark, 2001; Goldsmith, 2001; Onnis, Roberts & Chater, 2002).

The question of the learnability of a specific linguistic constraint (e.g. a constraint embodied in, say, the principles and parameters framework) can then be framed as follows: is the description length used by encoding the constraint offset by the saving in codelength achieved by encoding the data more precisely? If so, then this linguistic constraint can be learned from the corpus; otherwise, it is not learnable from the corpus. Thus, learners can add constraints to prune over-general models of the language, but only when adding these constraints leads to a shorter overall description of the linguistic input.

To make matters concrete, consider a simple constraint: that nouns and verbs must agree in number (singular vs. plural). Describing this constraint will take a certain amount of code – say 10s or 100s of bits (where a bit is the amount of information required to encode a binary symbol). To an approximation, this constraint halves the number of sentences in the language (thus, *the cow sings* and *the cows sing* are allowed, but **the cow sing* and **the*

cows sings are not). Using standard information theory, this means 1 bit is saved per sentence (Shannon, 1948). If a three-year-old has received a corpus of several million sentences/year, then this constraint would save several million bits; and so the constraint is clearly learnable. An interesting future project is to apply this approach more generally – thus providing a formal and implementation-independent analysis of learnability that is complementary to some of the specific computational models that MacWhinney describes.

The approach advocated by MacWhinney, and the results outlined in this commentary, aim to reframe a general LOGICAL problem of language acquisition as a series of empirical problems concerning the learnability of specific linguistic phenomena given the corpus (and other environmental) information available to the child. It may be hoped that further developments of the methods discussed by MacWhinney, and those described here, may help resolve the extent to which innate linguistic principles are required to explain human language acquisition.

REFERENCES

Barron, A., Rissanen, J. & Yu, B. (1998). The minimum description length principle in coding and modelling. *IEEE Transactions on Information Theory* **44**, 2743–60.

Brent, M. R. & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* **61**, 93–126.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review* **103**, 566–81.

Clark, R. (2001). Information theory, complexity, and linguistic descriptions. In S. Bertolo (ed.), *Parametric linguistics and learnability*. Cambridge: CUP.

Ellison, M. (1992). The machine learning of phonological structure. PhD thesis, University of Western Australia.

Gold, E. M. (1967). Language identification in the limit. *Information and Control* **16**, 447–74.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics* **27**, 153–98.

Grünwald, P. D. (1996). A minimum description length approach to grammar inference. In G. Scheler, S. Wermter & E. Riloff (eds), *Connectionist, statistical and symbolic approaches to learning for natural language processing*. New York: Springer.

Helm, P. van der (2000). Simplicity versus likelihood in visual perception: from surprisals to precisals. *Psychological Bulletin* **126**, 770–800.

Horning, J. J. (1969). A study of grammatical inference. Technical Report CS 139, Computer Science Department, Stanford University.

Li, M. & Vitányi, P. (1997). *An introduction to Kolmogorov complexity theory and its applications* (2nd edn). Berlin: Springer.

Onnis, L., Roberts, M. & Chater, N. (2002). Simplicity: a cure for overregularization in language acquisition. In L. R. Gleitman & A. K. Joshi (eds), *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal* **27**, 379–423 and 623–56.

Solomonoff, R. J. (1978). Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory* **24**, 422–32.