An Unambiguous Strategy for Learning Complex Linguistic Systems

Lisa Pearl, University of California, Irvine

Nov 26, 2007

University of Southern California Linguistics Colloquium

Language learning is a tricky business: the data are often noisy and the system generating the data doesn't necessarily have a transparent relationship with the observable data. Yet, children still seem to converge on the correct generalizations about the underlying systems of their native language.

While it can help to know the range of possible systems (in the form of parameters (Chomsky, 1981; Halle & Vergnaud, 1987) or constraints (Tesar & Smolensky, 2000)), this doesn't solve the problem of language learning by any means. Children must still choose from among this range given data that can be ambiguous and exception-filled. Learners might also employ a probabilistic learning strategy, so that learning is gradual and more robust to noise (Yang, 2002). Yet this, too, doesn't solve everything – the learner may still fall prey to exceptions if they are more numerous, and must still have some sort of process for dealing with ambiguous data.

Another strategy is to alter what data to learn *from*. Perhaps when they are deciding between generalizations, learners ignore data perceived as not very informative and focus instead on data perceived as maximally informative - that is, *unambiguous* data (Pearl & Weinberg, 2007; Lightfoot, 1999; Dresher, 1999; Fodor, 1998). A probabilistic learner that filters the data intake down to unambiguous data only may be able to succeed at learning complex systems, even if the data are highly noisy. Yet this, too, has its problems: for complex systems, do unambiguous data actually exist for individual values/constraints (Clark, 1994)?

In this study, I demonstrate the viability of data filtering for learning a realistic complex system from a realistic data set. The system is an instantiation of metrical phonology that has 9 interactive parameters (adapted from Dresher (1999)). The target system is English, and the input set is extrapolated from highly noisy English child-directed speech (CHILDES: MacWhinney, 2000). Given the non-trivial system to learn and the non-trivial data set to learn from, simulated learners can nonetheless converge on the correct generalizations for English. This is no small feat, and provides empirical validation for both the feasibility and sufficiency of an unambiguous data filtering bias.

REFERENCES:

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Clark, R. (1994). Kolmogorov complexity and the information content of parameters. *IRCS Report 94-17*. Institute for Research in Cognitive Science, University of Pennsylvania.

Dresher, E. (1999). Charting the learning path: Cues to parameter setting. *Linguistic Inquiry*, *30*, 27-67.

Fodor, J. D. (1998). Unambiguous Triggers. *Linguistic Inquiry, 29*, 1-36.

Halle, M. & Vergnaud, J-R. (1987). *An essay on stress.* Cambridge, MA: The MIT Press.

Lightfoot, D. (1999). *The Development of Language: Acquisition, Change, and Evolution.* Oxford: Blackwell.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pearl, L., & Weinberg, A. (2007). Input Filtering in Syntactic Acquisition: Answers from Language Change Modeling, *Language Learning and Development, 3(1),* 43-72.

Tesar, B. & Smolensky, P. (2000). *Learnability in Optimality Theory.* Cambridge, MA: The MIT Press.

Yang, C. (2002). *Knowledge and Learning in Natural Language.* Oxford: Oxford University Press.