Jack only learns from this data point, but Lily learns from that one, too.

Lisa Pearl
(in collaboration with Jeff Lidz)
University of Maryland
Center for Language Sciences: University of Rochester
May 14, 2007

Learning language is a tricky business: the system is complex, the data are often ambiguous, and the learner must frequently integrate data spanning multiple levels of representation. One potential solution is that learners filter the data used for learning (their data intake) down to a subset that is perceived as more informative about the underlying system, rather than using all available data. However, once filtering is invoked, we must investigate its feasibility, sufficiency, and necessity. What defines "informative" data and are there enough "informative" data in the available input? If the learner uses these data, does correct behavior result? Does incorrect behavior result without filtering?

Computational modeling is a valuable tool for addressing questions of data intake restriction, since they would be logistically (and ethically) difficult to explore with traditional experimental techniques. Moreover, computational modeling can use as boundary conditions the linguistic representations and the time course of acquisition that come from theoretical and experimental work. The computational modeling case study described here will be embedded in a framework that is applicable to a range of language learning problems. In addition, this framework combines discrete linguistic representations with probabilistic methods such as Bayesian updating which allows it to account for the gradualness and variation in learning that human children display.

In this talk, I will examine data intake filtering for learning English anaphoric *one* ("Jack only learns from this data point, but Lily learns from that *one*, too"), drawing on empirical data from experimental work by Lidz, Waxman, and Freedman (2003) as well as child-directed speech distributions from CHILDES (MacWhinney, 2000). Adult knowledge of anaphoric *one* has both a structural component (what the linguistic antecedent of *one* is) and a referential component (what *one* refers to in the world). Information for the learner is thus available from both the syntactic structure of the utterances containing anaphoric *one* and the situation in the world regarding what object *one* refers to. Moreover, the associated syntactic and semantic referent hypothesis spaces are linked: a linguistic antecedent (e.g. *data point*) will have a semantic referent in the world (e.g. DATA POINT), which allows knowledge coming from one source (e.g. syntactic structure) to affect the linked hypothesis space (e.g. predicted referent). Learners must integrate both sources of information into their linked hypothesis spaces that they then use for interpreting anaphoric *one*. The results from computational modeling suggest that data intake filtering is feasible, sufficient, and (perhaps surprisingly) necessary for successful acquisition of anaphoric *one*.

References:

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition, 89,* B65-B73.

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