Recent computer simulation studies of syntax acquisition necessarily rely on severely small, highly specified artificial language domains. Taking a different approach, we present a probabilistic framework that can predict learning behavior in a wide variety of linguistically interesting domains. Specifically, we abstract away from a syntactic characterization of the input sample encountered by the learner by formalizing a generally applicable notion of cross-language ambiguity. In this abstract, the focus is on domains approaching a realistic number of languages. This is important because, frequently, a tacit motivation for simulation studies (as versus probabilistic modeling) is that a learner’s performance scales more or less linearly as the size of the language domain grows – effective learning in a small, artificial domain should imply effective learning in the very large domain of human languages. An application of our framework to Gibson & Wexler’s TLA (1994) demonstrates that in some cases this motivation is misguided.

The TLA is a procedure for acquiring the settings of syntactic parameters. We address the question of feasibility – is acquisition possible within a reasonable amount of time and/or with a reasonable amount of work? Niyogi & Berwick (1996) and Turkel (1996) provide results from computer simulations run in the 3-parameter space originally constructed by G&W. More recently, Kohl (1999) completed a study of TLA performance in an expanded 12-parameter domain. Our investigation involves the building of several models to gauge the feasibility of TLA acquisition in more realistically sized parameter spaces. Due to space limitations, only two results are reported below.

Two standard assumptions are folded into the model: 1) sentences are arbitrarily drawn from the target language, and 2) at any given point, the learner entertains a current hypothesis which can change during the course of acquisition. Subsequently, we can assign \( \psi_i \): the probability of a successful parse of an arbitrarily encountered input sentence (generated by the target grammar), given that the learner’s current hypothesis is \( G_i \). The assignment of varying \( \psi_i \)’s determines the shape or distribution of cross-language ambiguity. By applying a standard non-linear optimization search to a constructed probabilistic (Markov) structure, the shape of ambiguity most conducive to efficient learning is determined. Performance is measured as the expected number of sentences consumed by the learner before convergence on the target grammar.

One result reveals that the TLA performs extremely well in strongly smooth domains – domains which exhibit a correlation between the similarity of grammars and the languages that are generated by them (Figure 1). However, with a homogeneous distribution of \( \psi \) across grammars, the number of input sentences required by the TLA rises exponentially with the number of parameters that need to be set; slightly worse than a ‘blind guess’ learner that picks grammars at random (Figure 2). Notably, the TLA is acutely sensitive to changes in the shape of ambiguity within the domain; a small change has a large impact on learning performance.

In summary: our framework is a sort of hybrid. It is formal, in that it abstracts away from specific characterizations of the input sample encountered by the learner, and empirical, in that various acquisition scenarios need to be tested in order to determine optimal distributions that can be generated by the formal analysis. The hybrid approach allows informative feasibility studies to be undertaken on learners operating in linguistically interesting spaces (e.g. smooth) of plausible size.