Putting the Emphasis on Unambiguous: The Feasibility of Data Filtering for Learning English Metrical Phonology

The linguistic system children acquire is complex and the available data are often noisy. One solution for converging on the correct system is that the learning mechanism is biased. This bias can be constraints on the hypothesis space, instantiated as a parametric system in domains like metrical phonology (Halle&Vergnaud, 1987). It can also be constraints on how children filter their data intake, such as only using data perceived as maximally informative (Pearl&Weinberg, 2007; Dresher, 1999; Lightfoot, 1999; Fodor, 1998). Yet, how feasible is learning from realistic data with either constraint type? Are there sufficient informative data for each parameter value in ambiguous, exception-filled input (Clark, 1994)? I will show that it is possible to learn a metrical phonology system instantiated as 9 interactive parameters (schematized in figure 1) in the highly noisy dataset of English using only unambiguous data. This supports the viability of both the parametric system and the unambiguous data filter.

Previous work has examined methods a learner could use to recognize unambiguous data in ambiguous input (cues: Dresher (1999), Lightfoot (1999); parsing: Sakas&Fodor (2001), Fodor (1998)) as well as investigated an unambiguous filter's feasibility, sufficiency, and necessity for learning a parametric system from realistic data distributions (Pearl&Weinberg, 2007). Yet, many of the parametric systems probed previously have been relatively small, involving only a few interactive parameters (Pearl&Weinberg, 2007; Yang, 2002; Niyogi&Berwick, 1996). Previous work examining larger parametric systems (Fodor&Sakas, 2004; Sakas, 2003; Sakas&Nishimoto, 2002) has often not implemented a representation of individual learning that simulates the gradualness and variation found in real language learners (as noted by Yang (2002)), nor incorporated frequency information into the data available to the learner.

The metrical phonology system presented here is complex, involving multiple interactive parameters that cause the order of parameter-setting to be crucial since it affects the learner's perception of incoming data. Specifically, data initially perceived as ambiguous may be later perceived as unambiguous, and data initially perceived as unambiguous may be later perceived as exceptions to the currently known system. Moreover, the learning model presented here is embedded in a learning framework that combines discrete representations with probabilistic learning in order to achieve the gradual learning that children exhibit. It also learns from data distributions based on estimates from 540505 words of child-directed speech (CHILDES, MacWhinney 2000).

Despite the non-trivial system and noisy data, I find that a simulated learner is able to identify sufficient unambiguous data to converge on the correct metrical phonology system of English. Though the two methods of unambiguous data identification (cues and parsing) require different constraints for success, both are able to encapsulate the necessary knowledge for successful acquisition in a small number of parameter-setting order constraints. These results support both the viability of the unambiguous data filter, contributing to knowledge of the learning mechanism, and the viability of learning a complex parametric system, contributing to theoretical work on metrical phonology. In addition, these results generate experimentally-testable predictions involving parameter-setting orders in children.

Word Count: 491

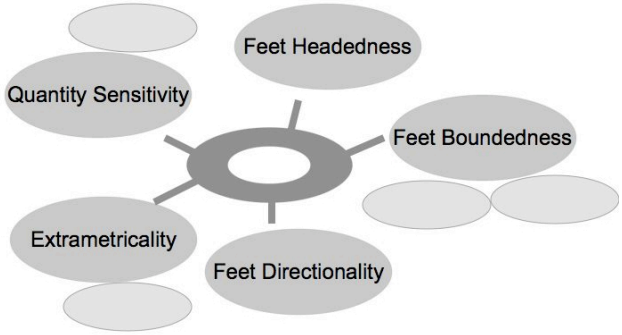

Figure 1. Schematization of interactive metrical phonology system: 5 main parameters and 4 sub-parameters.

REFERENCES:

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. (1998a). Unambiguous Triggers. *Linguistic Inquiry, 29*, 1-36.

Fodor & Sakas, 2004. Evaluating models of parameter setting. In A. Brugos, L. Micciulla and C. E. Smith (eds.) *BUCLD 28: Proceedings of the 28th Annual Boston University Conference on Language Development*. Cascadilla Press, Somerville, MA.

Halle, M. and J.-R. Vergnaud (1987) *An Essay on Stress*. Cambridge: 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.

Niyogi, P., & Berwick, R. (1996). A language learning model for finite parameter spaces. *Cognition, 61,* 161-193.

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

Sakas, W.G. (2003). A Word-Order Database for Testing Computational Models of Language Acquisition. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* 415-422.

Sakas, W.G. & Fodor, J.D. (2001). The structural triggers learner. In S. Bertolo (ed.) *Language Acquisition and Learnability*, Cambridge University Press, Cambridge, UK.

Sakas, W. & Nishimoto, E. (2002). Search, Structure, or Statistics? A Comparative Study of Memoryless Heuristics for Syntax Acquisition. Ms., CUNY: New York.

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