Some Limits on Statistical Learning

Rebecca Morley (morley@cogsci.jhu.edu)
William Badecker (badecker@cogsci.jhu.edu)
Paul Smolensky (smolensky@cogsci.jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles St. Baltimore, MD 21218

Introduction

A fairly large body of work within the artificial grammar paradigm has provided some reason to believe that statistical learning may be easy and rapid in both adult and infant subjects (e.g. Newport and Aslin, 2004; Wilson, 2003; Peperkamp et al., 2003). However, there are indications that domain and task may be very critical determiners of learnability. Factors such as degree of explicit knowledge in learning, strength of native language bias, and computational complexity of the data sample seem likely to play a role as well.

In the work presented here, subjects were exposed to a lexicon of novel words that exhibited a co-occurrence restriction of some type. Experiment 1 imposed a phonotactic constraint of within cluster voicing assimilation (VCCV word forms). Experiment 2 trained subjects on a (non-explicit) suffix alternation conditioned by the identity of the final stem consonant (CVCV-CV word forms). 2.2 was designed to test the hypothesis that the identity relation (CV<u>z</u>V-<u>z</u>i) may be necessary to bootstrap learning in this domain.

Methods

All trials consisted of a pre-test, training, and post-test portion. The pre-test and post-test task was a two-alternative forced choice paradigm which asked subjects to choose which of two auditorily presented tokens was a word in the made-up exposure language (one alternative was grammatical and one ungrammatical according to the particular alternation). Exposure consisted of approximately 20 minutes of auditorily presented word tokens from one of the following data distributions.

Data Distributions

1.1

VvbV	VbvV	VgbV	VzvV
VvgV	VbgV	VgvV	VzbV
VvzV	VbzV	VgzV	VzgV
VkfV	VfkV	VpkV	VskV
VkpV	VfpV	VpfV	VsfV
VksV	VfsV	VpsV	VspV

1.2

VvzV	VbzV	VgzV
VksV	VfsV	VpsV

2.1

CVzV-zi	CVsV-si	CVfV-si	CVpV-si
CVvV-zi	CVbV-si	CVkV-si	CVgV-si

2.2

CVgV-zi	CVsV-si	CVfV-si	CVpV-si
CVvV-zi	CVbV-si	CVkV-si	

Results & Conclusions

Subjects failed to learn in experiments 1.1,1.2, and 2.2, and demonstrated successful learning in experiment 2.1 (for both old and new forms). These results suggest that learning of novel linguistic distributions based on co-occurrence statistics may be more difficult to induce than previously supposed: that the computational device may be quite limited in the number of units it can keep track of, that a highly salient identity relation may be necessary to focus attention on potential structure, etc. Further work will be needed to fully disentangle the contributions of each of these factors.

Acknowledgments

This work was supported by an NSF IGERT training grant to the Department, and a Department of Education Javits Fellowship.

References

Newport,E.L. & Aslin, R.N. (2004). Learning at a Distance I. Statistical Learning of Non-adjacent Dependencies. *Cognitive Psychology*, 48, 127-162.

Peperkamp, S.,Pettinato, M., et al. (2003). Allophonic variation and the acquisition of phoneme categories. *BUCLD 27, 650-661.*

Wilson, C. (2003). Experimental Investigation of Phonological Naturalness. *West Coast Conf. Formal Ling. 22,* 101-114.