Speaker: **Gaja Jarosz**
University of Massachusetts, Amherst

Title: **Maximum Likelihood Learning of Phonology**

Date: Monday, February 05, 2007
Time: 4:00pm
Place: Rosenfeld Hall (corner Temple & Grove Streets)

Abstract:

Learning phonology involves learning a grammar, a lexicon of underlying forms, as well as the inaudible structure assigned by the grammar to overt phonological forms. Because of the complexity of this task, most work on phonological learning has focused on some sub-problem of phonological learning.

For example, the most well-know work on learnability within Optimality Theory (Tesar and Smolensky 1998, Boersma and Hayes 2001) focuses on the task of learning a grammar given the correct underlying forms and structured surface forms. In this talk, I present MLG, a theory of phonological learning that tackles the full problem of learning a grammar and lexicon, given unstructured phonological forms. MLG draws on insights from phonological theory, statistical language learning, and language acquisition to address the various challenges posed by this modeling task.

One major challenge is the learning of grammars with generalizing capacity that are simultaneously restrictive, two conflicting preferences. The proposed solution, Maximum Likelihood Learning of Lexicons and Grammars (MLG), combines a probabilistic formulation of Optimality Theory (Prince and Smolensky 1993/2004) with statistical learning via likelihood maximization. The central premise of the proposed theory of phonological learning is that the correct grammar and lexicon combination makes the overt forms most expected, or likely.

The generalizing capacity of grammars is attributed to phonological theory, which defines the space of possible languages. Because of implicational markedness universals and phonological features and representations, learnable languages are general, and arbitrary gaps cannot be learned. The identification of restrictive grammars is the consequence of maximum likelihood learning in conjunction with explicit reliance on richness of the base, a fundamental Optimality Theoretic principle. Moreover, the particular instantiation of a rich base draws on findings from phonological acquisition suggesting that children divide the phonological

learning problem into two subproblems: the learning of a restrictive phonotactic grammar and the learning of alternations.

A second major challenge involves the identification of the optimal grammar and lexicon combination in the face of ambiguity and variation. Maximum likelihood learning, via the EMGL algorithm, enables the successful identification of the correct solution by gradually converging on the grammar and lexicon combination that best explains the observed data. EMGL is a variant of the well-known Expectation-Maximization algorithm. I will show how this procedure successfully learns general, restrictive grammars and correct lexicons in a variety of language systems involving different kinds of hidden structure: syllable structure, yer vowels, voicing neutralization, and free variation.

Finally, I discuss the predictions of MLG for the process of acquisition. The proposed learning theory builds a foundation for the computational modeling of child phonological acquisition, accounting for the end states of two stages of acquisition, phonotactic learning and morphophonemic learning, as well as the gradual transition between these stages. Markedness is predicted to be the primary constraint on possible acquisition paths, with frequency playing a secondary role. I discuss stage-by-stage predictions for the acquisition of onsets of varying sonority and complexity in Polish based on observed frequencies of various onsets in Polish.