

STANFORD UNIVERSITY  
DEPARTMENT OF STATISTICS  
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, October 17, 2000  
Sequoia Hall Rm. 200  
(Cookies at 3:45 in 1st Floor Lounge)

*Thorsten Brants*  
*Xerox PARC*

**Part-of-Speech Tagging with Hidden Markov Models**

Hidden Markov models represent random sequences of states, and each state randomly selects a symbol that is emitted. While at first sight a doubly random process is completely unrelated to the production and perception of speech and language, HMMs nevertheless turn out to be very successful for a large number of tasks in this domain. We will focus on part-of-speech tagging, i.e. the unique annotation of a word with a syntactic category. The advantage of HMMs for this task are:

- HMMs yield very high accuracies, outperforming most other techniques;
- they achieve high processing speeds;
- the model learns from existing data;
- HMMs assign probabilities for rankings and reliability estimates.

This talk will explain the representation of the tagging task as an HMM, the generation of an HMM from textual data, handling of sparse data, and efficient processing of HMMs.