Computation, Information, and Intelligence (ENGRI/CS/INFO/COGST 172), Spring 2007
4/11/07: **Lecture 32 aid — learning bigram models**

Topics: Unsupervised learning of a very restricted class of weighted (or probabilistic) grammars; language-model smoothing.

I. More on garden-path sentences Some observations we can make from the garden-path phenomenon:

- Failure to parse yields a failure to understand.
- Humans engage in on-line processing, constructing hypotheses as we go.
- Humans discard, or at any rate do not consider, all possible correct hypotheses.

II. Sentence-ranking example A classic from the speech-recognition literature.

1. It's hard to recognize speech.
2. It's hard to wreck a nice beach.

III. Bigram CFGs [1] A bigram CFG would take the following form:

- Terminals: $w_1, w_2, \ldots, w_m$, the "real words", plus a special "end of sentence" terminal w_{m+1} that is inserted at the end of every sentence and that appears nowhere else in any sentence.
- Nonterminals: $S, V_1, V_2, \ldots, V_{m+1}$
- Start symbol: S
- Rewrite rules: all rewrite rules of the form
  1. $V_i \rightarrow w_i V_j$,
  2. $S \rightarrow V_i$

  where $1 \leq i \leq m$, $1 \leq j \leq m+1$, plus the rule $V_{m+1} \rightarrow w_{m+1}$.

IV. The poverty of the stimulus The classic example, due to Noam Chomsky:

1. Colorless green ideas sleep furiously.
2. Furiously sleep ideas green colorless.

V. Interpolation smoothing For i between 1 and m inclusive, set the probability of a rule $V_i \rightarrow w_i V_j$ (which, in our case, corresponds to the probability that if word w_i occurs then word w_j follows it) to

$$\lambda \frac{\#(w_i w_j)}{\sum_k \#(w_i w_k)} + (1-\lambda) \frac{\#(w_j)}{\sum_k \#(w_k)}$$

where the interpolation parameter λ (pronounced "lambda") is between 0 and 1 (usually non-inclusive).

[1] This definition improves on that given in the previous lecture aid (which we didn't get to anyway) in terms of probabilistic estimation.