Computation, Information, and Intelligence (ENGRI/CS/INFO/COGST 172), Spring 2007
4/4/07: **Lecture 29 aid — Finish Earley parsing**

Topics: Earley's algorithm for CFG parsing, continued.

Announcements: HW4B mean (μ): 1.6 out of 2, standard deviation (σ) 0.6; 4C μ: 4.5 out of 5, $\sigma = 0.7$.

I. Earley's algorithm: parser actions Suppose $\underline{x} = x_1 x_2 \ldots x_n$ is the n-word sentence being parsed.[1]

1. *Scan* (advance the dot over a terminal that matches the next word in the sentence):[2]

   Given $(X \rightarrow \alpha \bullet y\beta, i, j)$ where $y = x_{j+1}$, add state $(X \rightarrow \alpha x_{j+1} \bullet \beta, i, j+1)$.

2. *Complete* (advance the dot over a nonterminal that accounts for the next subsequence of words in the sentence):

   Given $(Y \rightarrow \alpha \bullet, j, k)$ and $(X \rightarrow \beta \bullet Y\gamma, i, j-1)$, add state $(X \rightarrow \beta Y \bullet \gamma, i, k)$.

3. *Predict* (guess how to account for the next part of the sentence):

   Given $(X \rightarrow \alpha \bullet Y\beta, i, j)$ and rewrite rule $Y \rightarrow \gamma$, add state $(Y \rightarrow \bullet\, \gamma, j+1, j)$.

   Note the special treatment of what are usually endpoint indicators, meant to remind us that we don't know how far into the sentence that γ will "cover".

II. Earley's algorithm This algorithm is guaranteed to terminate.

1. Start with the special state ($\rightarrow\bullet$S, 1, 0).
2. Perform all possible predictions.
3. Scan.
4. Perform all possible completions.
5. Return to step 2 unless nothing changed from the previous round.

III. Self-check Run through the example execution given on last lecture's aid. See that you can reconstruct the partial parses that are represented by the parse states given. What you should observe is that the hypothesis that "the sea" forms the entirety of the subject of the sentence gets "stuck" once $x_3 =$ "turtle" is observed.

---

[1] The information presented here is completely equivalent to that given on the last lecture aid; there are just a few adjustments of notation. For instance, the change of notation for the sentence being parsed (i.e., from $w_1 w_2 \ldots w_n$) is meant to prevent confusion with the notation for unique terms in the information-retrieval unit of the course. The point is that sentences can repeat words, e.g., we might have $x_1 = x_2 =$ "never" in "never never do that again", but in the IR unit we had been using w_1 and w_2 to indicate two distinct terms in the vocabulary.

[2] A few remarks about Greek letters. First, a pronunciation guide:

α	alpha
β	beta
γ	gamma

Why do these come up now? Well, by convention, lowercase is used to indicate terminals and uppercase is used to indicate nonterminals, so the convention is to resort to an entirely different alphabet to indicate mixtures of the two. (If you have any suggestions for less forbidding and yet equally concise notation, I would be happy to entertain suggestions. Last year's crew seemed to find the Greek somewhat comical, or something.)