The anatomy and interpretations of states in the Earley algorithm

A state consists of four parts:

A context-free rule
An integer: the "dot"
An integer: ***i***
An integer: ***j***

Together these four ingredients represent a hypothesis of a parse for a substring of the input:

The rule represents the structure that is hypothesized for the substring

The dot represents how much of the right-hand side of the rule has been confirmed, so far, by comparing with the input. What is to the left of the dot has already been found in the input. The symbol to the right of the dot (if any) is the prediction as to what comes next in the input. (In the example states below the dot information is represented by the position of the period character in the right-hand side string.)

The first integer, ***i***, represents the beginning point of the substring.

The second integer, ***j***, represents the point of progress so far in the input string.

Examples:

S ® NP .VP 0, 4
This state represents the hypothesis that beginning at position 0 (the left edge) of the input there is a substring parsable as ***S***, with the structure [***NP VP***]. So far, between 0 and 4, a ***NP*** has been found. The state predicts that a ***VP*** will come next, starting at position 4 of the input string.

S ® NP VP. 3, 8
This state represents the hypothesis that beginning at position 3 (after the third word) of the input there is a substring parsable as ***S ® NP VP***. The fact that the dot is at the end of the right-hand side represents the fact that both the ***NP*** and ***VP*** have been found. Therefore, the substring between positions 3 and 8 of the input has been successfully parsed as ***S***. This state makes no prediction as to what kind of phrase begins at position 8.

S ® .NP VP 4, 4
This state represents the hypothesis that beginning at position 4 of the input (the fifth word), there is a substring parsable as ***S ® NP VP***. So far neither the ***NP*** nor the ***VP*** has been found. So the prediction is that the next constituent, an ***NP***, begins at position 4 of the input.

S ® .NP VP 0,0

This state represents the hypothesis that beginning at position 0 of the input (the first word) there is a substring parsable as ***S ® NP VP*** . So far neither the ***NP*** nor the ***VP*** has been found. So the prediction is that the next constituent, an ***NP***, begins at the first word of the input.

Notice that whenever the dot is at the beginning of the right-hand side, the two integers ***i*** and ***j*** will be identical, as in the last two states above. When the dot is in some other position, the two integers will be different, with the second always being more than the first. In that case the second integer reflects the fact that some progress has been made toward a successful parse of the substring.

F ® .S 0,0

This is the so-called “seed state”. It consists of forming a state by making a new rule (not in the grammar) with a dummy symbol (***F***) as its left-hand side, and on the right-hand side the initial symbol of the grammar, with the dot at the beginning of the left-hand side, and integers ***0,0***. The function of the seed state is to give the algorithm something to work on as parsing begins.