Deterministic Recognizers for CFL's

As an application of Chomsky Normal Form, we can obtain a deterministic algorithm for recognizing a context-free language.

Def. A *recognizer* for a context-free language L is an algorithm that inputs a string w and determines whether or not $w \in L$.

Suppose $L = L(G)$, where G is in Chomsky Normal Form. Given $w = w_1 w_2 \dots w_n$, define

$$D(i, l, A) \;=\; \text{true iff } A \stackrel{*}{\Rightarrow} w_i, w_{i+1}, \dots w_{i+l-1}.$$

for $A \in V$, $1 \leq i \leq n$ and $1 \leq l \leq n$.

Note that:

- $D(i, l, A)$ is true if and only if either
  - G has a rule $A \to a$ with $w_i = a$, or
  - G has a rule $A \to BC$, and $D(i, k, B)$ and $D(i+k, l-k, C)$ are both true, for some k with $1 \leq k < l$.
- $w \in L$ if and only if either
  - $w = \epsilon$ and $S \to \epsilon$, or
  - $D(1, |w|, S)$ is true.

The Cocke-Younger-Kasami (CYK) Algorithm

The CYK algorithm determines whether a given $w = w_1 w_2 \dots w_n$ ($w \neq \epsilon$) is in L, by computing the boolean values $D(i, l, A)$ for $A \in V$, $1 \leq i \leq n$, and $1 \leq l \leq n$.

- **Initialization:**

  set all $D(i, l, A)$ to false
  for i from 1 to n
      for each rule $A \to a$
          if $w_i = a$ then set $D(i, 1, A)$ to true.

- **Main Loop:**

  for l from 2 to n (length)
  for i from 1 to $n-(l-1)$ (start posn.)
  for k from 1 to $i-1$ (partition)
  for each rule $A \rightarrow BC$
  if $D(i,l,B)$ and $D(i+k,l-k,C)$ are true
  then set $D(i,l,A)$ to true.

- **Result:** If $D(1,n,S)$ is true then output YES else output NO.

Notes on the CYK Algorithm

- The worst-case running time is $O(n^3 \cdot |G|)$, where $|G|$ is the size (*e.g.* # of symbols to write it down) of G.
- The $O(n^3 \cdot |G|)$ worst-case running time is similar to other general CFL recognition algorithms (*e.g.* Earley's algorithm).
- Valiant (1975) showed how *fast boolean matrix multiplication* can be used to compute the $D(i, l, A)$.
- The best current matrix multiplication algorithm (Alman-Williams, 2020) gives a CFL recognizer that runs in time $O(n^{2.3728596})$ (but is not practical).

Programming language parsers use linear-time algorithms that only work for *deterministic* CFL's.