One main reason why a CNF is useful

Main result .

The result we showed in class was basically establishing the following. Given a grammar in CNF, we are guaranteed that a derivation for a string of length n has to be of length $2n-1$.

Here then is the main idea of the proof.

We proceed by first recalling what the permitted type of rules of a CNF are

- A variable going to a single terminal $A \to a$
- A variable going to exactly 2 variables $A \to BC$, where either one of B or C is allowed to be the same as A as well.
- $S \to \varepsilon$
- S can appear only on the left side of a rule

Given these restrictions consider the following questions

1. How long is a derivation of a string with just 1 character? That is trivial to answer because the only possible first step in this case is $S \to a$.
2. How long is a derivation of a string with just 2 characters? We know that we want to get from S to something with 2 terminals like ab.

   But this immediately tells us that our first step cannot possibly be something $S \to a$ (a terminal does not allow us to move forward in any way).

   But if you cannot allow for $S \to a$ the only other possibility is a rule of the form $S \to AB$

   Now consider how to get from AB to two terminal symbols ab. You will see that this will require 2 steps. Putting it all together, we have a 3 step process to get to a 2 character string.

   The standard technique when you want to claim a pattern holds is . . . induction.

One subtlety that I did not make explicit in class is that to prove this by induction you do want to make a stronger claim which is

Claim

$A \stackrel{*}{\Longrightarrow} w_1 w_2 \ldots w_n$ where each w_i is a single character takes $2n-1$ steps. That is, the property is not just restricted to the start symbol.

We have already shown some base cases.

The induction hypothesis (we will use strong induction) says that any variable going to any string of size k less than n, takes $2k-1$ steps.

Now consider what the first step for the derivation of an n character string could possibly be!

Anything of the form $S \to a$ is not possible, so it must be something of the form $S \to AB$.

That immediately means that $A \stackrel{*}{\Longrightarrow} w_1 w_2 \ldots w_k$ for some k smaller than n and $B \stackrel{*}{\Longrightarrow} w_{k+1} w_{k+2} \ldots w_n$.

Applying the induction hypothesis, we get the total number of steps as $2k-1+2(n-k)-1+1$, which sums up to $2n-1$.