An example of a formal proof that a CFG generates exactly a given set of words

Prakash Panangaden

October 14, 2018

Consider the grammar

$$S \to aSb|bSa|SS|\varepsilon.$$

What set is generated by this grammar? We need to say "the language consists of all string of such and such type." Then we need to *prove* that every string generated by the grammar has the property that we claimed and *also* that every string with that property can be generated by the grammar. Thus there will be *two* proofs.

This grammar, which we will call G, generates the set L of words with equal numbers of a's and b's.

To prove this, we must show that any word generated by the grammar is in L, ie has equal numbers of a's and b's, and conversely that any word in L is generated by the grammar.

For the first, we can simply analyse the productions. There are only two productions in which letters are added; each adds a single a and a single b. So, whenever letters are added, they are added in equal numbers, and it follows that any string generated by G has equal numbers of as and bs, *i.e.* is in L.

The second is a little trickier. We proceed by induction on the number n of a's — or equivalently, the number of b's — in the strings of L. For the base of our induction, the string in L with no a's is the empty string ε, and this is generated by the grammar. Now assume that all strings in L with no more than n a's are generated by G; we must now show that any string w in L with $n+1$ a's is also generated.

Suppose w begins and ends with the same letter — a, say — ie, $w = aw'a$.

Then w' contains two more b's than a's, and so some proper prefix of w' must contain exactly one more b than it does a's. That is, $w' = xy$ with x having one more b than a's; y must have the same property, as w' in total has two more b's than a's. So now $w = ax \cdot ya$, and ax and ya each have equal numbers of a's as b's and certainly no more than n a's, so each is generated by G. Then w is generated by the production $S \to SS$.

Suppose instead that w begins and ends with different letters — $w = aw'b$, say. Then w' is in L and so, by the induction hypothesis, w' is generated by G. Then w can be generated using the rule $S \to aSb$ by generating w' with the middle S. The situation for $w = bw'a$ is similar.

Thus any string in L can be generated by G, and so G generates exactly the set of words with equal numbers of a's and b's, as claimed.