1 Closure Properties of Context-Free Languages

We show that context-free languages are closed under union, concatenation, and Kleene star. Suppose $G_1 = (V_1, \Sigma_1, R_1, S_1)$ and $G_2 = (V_2, \Sigma_2, R_2, S_2)$.

Example: For G_1 we have

$$S_1 \to aS_1b$$

$$S_1 \to \epsilon.$$

For G_2 we have

$$S_2 \to cS_2d$$

$$S_2 \to \epsilon.$$

Then $L(G_1) = \{a^nb^n : n \geq 0\}$. Also, $L(G_2) = \{c^nd^n : n \geq 0\}$.

1.1 Union

$G = (V_1 \cup V_2 \cup \{S\}, \Sigma_1 \cup \Sigma_2, R, S)$ where $R = R_1 \cup R_2 \cup \{S \to S_1, S \to S_2\}$ and S is a new symbol.

Then $L(G) = L(G_1) \cup L(G_2)$.

Example:

$$S_1 \to aS_1b$$

$$S_1 \to \epsilon.$$

$$S_2 \to cS_2d$$

$$S_2 \to \epsilon.$$

$$S \to S_1$$

$$S \to S_2$$

Then $L(G) = \{a^nb^n : n \geq 0\} \cup \{c^nd^n : n \geq 0\}$.

1.2 Concatenation

$G = (V_1 \cup V_2 \cup \{S\}, \Sigma_1 \cup \Sigma_2, R, S)$ where $R = R_1 \cup R_2 \cup \{S \to S_1S_2\}$ and S is a new symbol.

Example:

$$S_1 \to aS_1b$$

$$S_1 \to \epsilon.$$

$$S_2 \to cS_2d$$

$$S_2 \to \epsilon.$$

$$S \to S_1S_2$$

Then $L(G) = \{a^mb^mc^nd^n : m, n \geq 0\}$.

1.3 Kleene star

$G = (V_1 \cup \{S\}, \Sigma_1, R, S)$ where $R = R_1 \cup \{S \to \epsilon, S \to SS_1\}$ and S is a new symbol.

Example:

$$S_1 \to aS_1b$$

$$S_1 \to \epsilon.$$

$$S \to \epsilon$$

$$S \to SS_1$$

Then $L(G) = \{a^nb^n : n \geq 0\}^*$.

Do some sample derivations.

1.4 Non-closure properties

Context-free languages are not closed under intersection or complement. This will be shown later.

1.5 Intersection with a regular language

The intersection of a context-free language and a regular language is context-free (Theorem 3.5.2). The idea of the proof is to simulate a push-down automaton and a finite state automaton in parallel and only accept if both machines accept.

- Using this result one can show for example that the set of strings having equal numbers of a and b but no substring of the form *abaa* or *babb* is context-free; this would be very difficult to do using grammars.
- As another example, $\{a^n b^n : n \geq 0\} \cap \{w \in \{a,b\}^* : |w| \text{ is divisible by } 3\}$ is context-free.

2 Showing languages are not context-free

This section will give formal methods to show that languages are not context-free. It will also help your intuition so that you will usually be able to tell right away whether or not a language is context-free.

- To show that a language L is not context-free, it is necessary to find a property P that all context-free languages have, and then show that L does not have property P.
- For context-free languages, the property P is a pumping property, similar to that for regular languages.

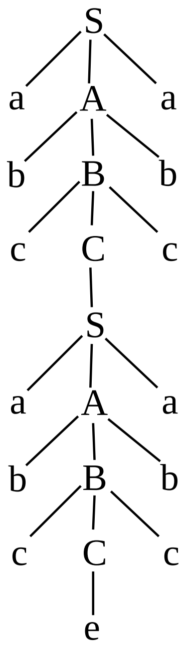
First we illustrate this property by an example.

2.1 Example of a property of all context-free languages

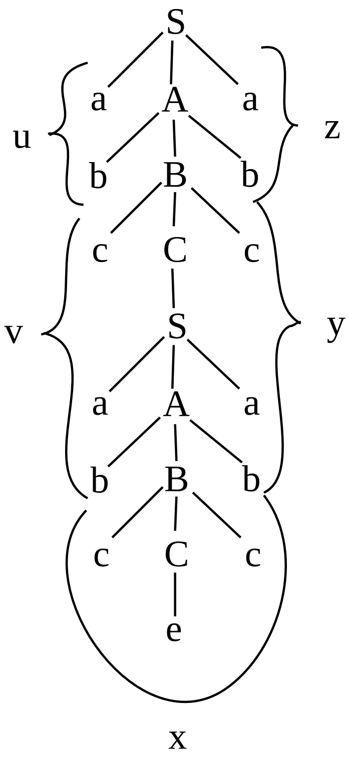
Suppose a context-free grammar G is (V, Σ, R, S) where $\Sigma = \{a, b, c\}$, $V = \{S, A, B, C\}$, and R consists of the following rules:

$$\begin{array}{ll} S \to aAa & C \to S \\ A \to bBb & C \to \epsilon \\ B \to cCc & \end{array}$$

Then we have the following parse tree for the string *abcabccbacba* in $L(G)$:



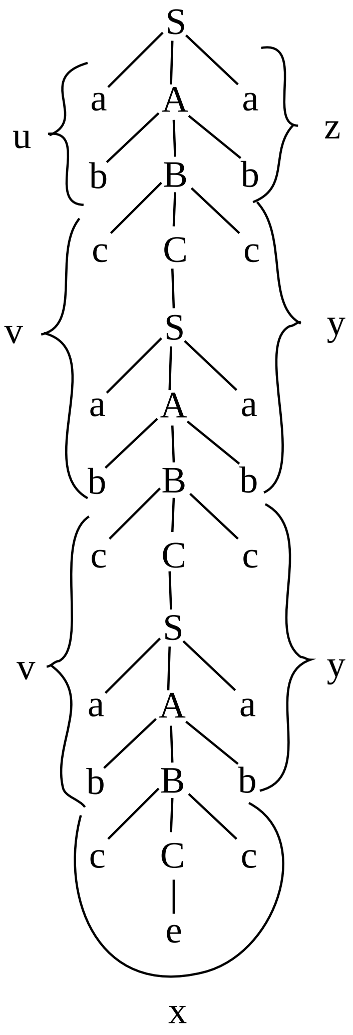
- Note that if a string is sufficiently long, a parse tree for the string will be very large, so it will have at least one very long path.
- This path will have some nonterminal appearing twice, just as the B (and other nonterminals) appear twice in this example.
- Now, using these two occurrences of B on the path, we can separate the string into substrings u, v, x, y, z as follows:



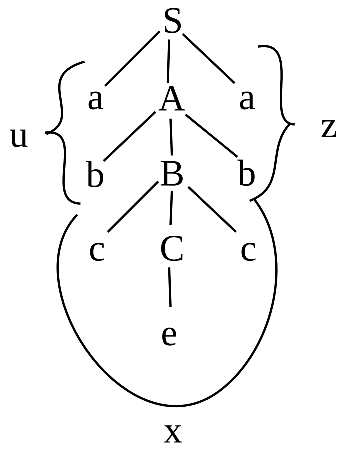
The string is divided according to what can be derived from the two occurrences of B in the parse tree.

- Thus $u = ab$, $v = cab$, $x = cc$, $y = bac$, and $z = ba$.
- So the string $abcabccbacba$ can be expressed as $uvxyz$ in this way: $(ab)(cab)(cc)(bac)(ba)$.

Now, note that the portion of the tree between the two B's can be duplicated, like this:



This gives the string $(ab)(cab)(cab)(cc)(bac)(bac)(ba)$, that is, $uvvxyyz$, or, uv^2xy^2z. In the same way, the portion of the parse tree between the two occurrences of B can be deleted, like this:



This gives the string $(ab)(cc)(ba)$, that is, uxz, or, uv^0xy^0z. In the same way, one can obtain uv^ixy^iz for any $i \geq 0$. Note that v and y are pumped at the same time.

Thus we have

$$S \Rightarrow^* abBba$$

$$B \Rightarrow^* cabBbac;$$

$$B \Rightarrow^* cc.$$

So we can write this as

$$S \Rightarrow^* uBz$$

$$B \Rightarrow^* vBy$$

$$B \Rightarrow^* x.$$

Note now that

$$B \Rightarrow^* vBy \Rightarrow^* vvByy \Rightarrow^* vvvByyy$$

et cetera, so in general

$$B \Rightarrow^* v^i B y^i$$

for all i. Thus we have

$$S \Rightarrow^* uBz \Rightarrow^* uxz,$$

$$S \Rightarrow^* uBz \Rightarrow^* uvByz \Rightarrow^* uvxyz,$$

$$S \Rightarrow^* uBz \Rightarrow^* uvByz \Rightarrow^* uvvByyz \Rightarrow^* uvvxyyz,$$

and in a similar way we have $S \Rightarrow^* uv^ixy^iz$ for all $i \geq 0$.

2.2 General property that all context-free languages have

In general, for any context-free language, large enough strings will have parse trees with long paths so that some nonterminal appears twice on the path. We can repeat the above argument, then, so we have the following property (from handout 6):

> If L is a context-free language, then there is an integer N such that any string $w \in L$ of length larger than N can be written as $uvxyz$ such that ($v \neq e$ or $y \neq e$) and $uv^ixy^iz \in L$ for all $i \geq 0$.

2.3 Using this general property to show languages are not context-free

Thus to show that a language is not context-free it is necessary to show that it does not have this property; this yields the following result, also from handout 6:

> If L is a language and

- for all integers N,
- there is a string $w \in L$ of length greater than N such that
- for all ways of writing w as $uvxyz$ with ($v \neq e$ or $y \neq e$),
- there is an i such that
- uv^ixy^iz is not in L,

then L is *not* context-free.

This can be used to devise a game to show that a language is not context-free, as illustrated on handout 6. Note that these methods cannot show that a language is context-free; only that a language is *not* context-free.

2.4 Showing that $\{a^nb^nc^n : n \geq 0\}$ is not context free

- As an example, using these methods, one can show that the language $\{a^nb^nc^n : n \geq 0\}$ is not context-free.
- For this, let $a^Nb^Nc^N$ be the string in L of length greater than N.
- Then we have to show that for all ways of writing $a^Nb^Nc^N$ as $uvxyz$ with ($v \neq e$ or $y \neq e$), there is an i (namely $i = 2$) such that uv^ixy^iz (namely $uvvxyyz$) is not in L.

To do this, there are two cases.

1. vy contains occurrences of all three symbols a, b, and c.

   Then v or y has to contain two different symbols, so $vvyy$ has a b before an a, a c before a b, or a c before an a, so the letters are out of order and $uvvxyyz$ is not in L.

2. vy contains occurrences of only two of the three symbols.

   In this case $uvvxyyz$ has unequal numbers of a, b, and c, so $uvvxyyz$ is not in L.

2.5 Example for the proof

Here is an example to illustrate the proof.

- Suppose N is 5, then $a^N b^N c^N$ is $aaaaabbbbbccccc$ or $a^5 b^5 c^5$.
- Suppose $u = aa$, $v = aaabb$, $x = bb$, $y = bccc$, and $z = cc$.
- Thus $uvxyz$ is $(aa)(aaabb)(bb)(bccc)(cc)$.
- Then v and y together have all three symbols, and v has both a and b.
- Then uv^2xy^2z is $(aa)(aaabb)^2(bb)(bccc)^2(cc)$, or, $(aa)(aaabb)(aaabb)(bb)(bccc)(bccc)(cc)$.
- This has a b before an a and is therefore not in L.

Now suppose that together v and y have only two of the three symbols. In particular, suppose that $u = aa$, $v = aaa$, $w = bbbbb$, $y = ccc$, and $z = cc$.

- Thus $uvxyz = (aa)(aaa)(bbbbb)(ccc)(cc)$.
- Now, v and y together have only a and c, no b.
- Then uv^2xy^2z is $(aa)(aaa)^2(bbbbb)(ccc)^2(cc)$.
- This string is $(aa)(aaa)(aaa)(bbbbb)(ccc)(ccc)(cc)$ which has 8 a, 5 b, and 8 c.
- The number of a and c in uv^2xy^2z has increased over that of $uvxyz$, but the number of b has not.
- So the number of a, b, and c is not the same in uv^2xy^2z, so uv^2xy^2z is not in L.

This proof can also be done using a game, as before.

2.6 Another example

Using this same approach, one can show that $\{a^p : p \text{ is prime } \}$ is not context-free.

2.7 Intersecting with a regular set

Now consider the language $L = \{w \in \{a, b, c\}^* : w$ has the same number of a, b, and $c\}$. Let R be $\mathcal{L}(a^*b^*c^*)$.

- Then $L \cap R = \{a^n b^n c^n : n \geq 0\}$ which we just showed is not context-free.
- If L were context-free then $L \cap R$ would be context-free as well, because context-free languages are closed under intersection with a regular set.
- Therefore L is not context-free.

This shows how one can sometimes use intersection with a regular language to show that a language is not context-free.

3 Non-closure properties of context-free languages

Theorem 3.1 (3.5.4) *Context-free languages are not closed under intersection or complement.*

Proof: Let $L_1 = \{a^m b^m c^n : m, n \geq 0\}$ and let $L_2 = \{a^m b^n c^n : m, n \geq 0\}$.

- Then both L_1 and L_2 are context-free.
- However, their intersection $L_1 \cap L_2$ is $\{a^n b^n c^n : n \geq 0\}$ which is not context-free.

For complement, note that $L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}}$ where $\overline{L}$ is the complement of L.

- If context-free languages were closed under complement, they would also be closed under intersection.
- Therefore context-free languages are not closed under complementation because they are not closed under intersection.