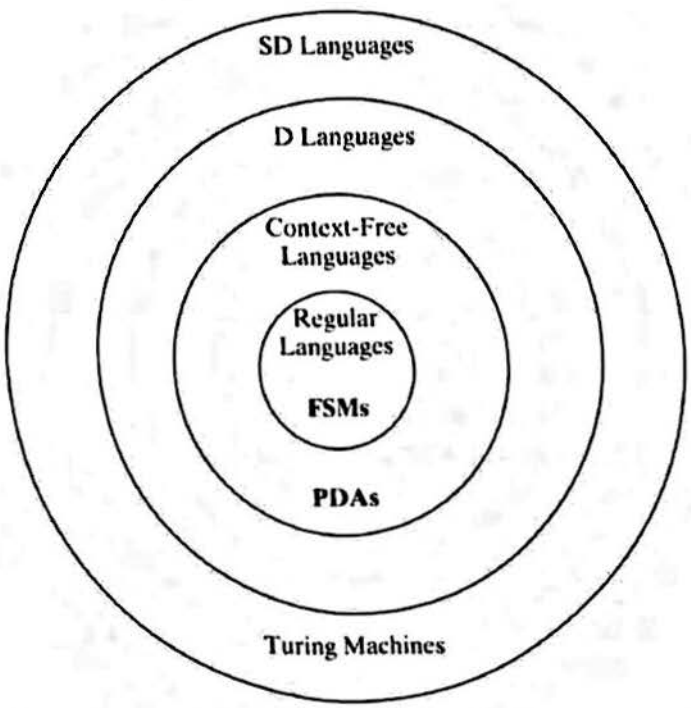
PART III

CONTEXT-FREE LANGUAGES AND PUSHDOWN AUTOMATA

In this section, we move out one level and explore the class of context-free languages.

This class is important. For most programming languages, the set of syntactically legal statements is (except possibly for type checking) a context-free language. The set of well-formed Boolean queries is a context-free language. A great deal of the syntax of English can be described in the context-free framework that we are about to discuss. To describe these languages, we need more power than the regular language definition allows. For example, to describe both programming language statements and Boolean queries requires the ability to specify that parentheses be balanced. Yet we showed in Section 8.4 that it is not possible to define a regular language that contains exactly the set of strings of balanced parentheses.



We will begin our discussion of the context-free languages by defining a grammatical formalism that can be used to describe every language in the class (which, by the way, does include the language of balanced parentheses). Then, in Chapter 12, we will return to the question of defining machines that can accept strings in the language. At that point, we'll see that the pushdown automaton, an NDFSM augmented with a single stack, can accept

exactly the class of context-free languages that we are about to describe. In Chapter 13, we will see that the formalisms that we have presented stop short of the full power that is provided by a more general computational model. So we'll see that there are straightforward languages that are not context-free. But, because of the restrictions that the context-free formalism imposes, it will turn out to be possible to define algorithms that perform at least the most basic operations on context-free languages, including deciding whether a string is in a language. We'll summarize those algorithms in Chapters 14 and 15.

The theory that we are about to present for the context-free languages is not as straightforward and elegant as the one that we have just described for the regular languages. We'll see, for example, that there doesn't exist an algorithm that compares two pushdown automata to see if they are equivalent. Given an arbitrary context-free grammar G, there doesn't exist a linear-time algorithm that decides whether a string w is an element of $L(G)$. But there does exist such an algorithm if we restrict our attention to a useful subset of the context-free languages. The context-free languages are not closed under many common operations like intersection and complement.

On the other hand, because the class of context-free languages includes most programming languages, query languages, and a host of other languages that we use daily to communicate with computers, it is worth taking the time to work through the theory that is presented here, even though it is less clear than the one we were able to build in Part II.

CHAPTER 11

Context-Free Grammars

We saw, in our discussion of the regular languages in Part II, that there are substantial advantages to using descriptive frameworks (in that case, FSMs, regular expressions, and regular grammars) that offer less power and flexibility than a general purpose programming language provides. Because the frameworks were restrictive, we were able to describe a large class of useful operations that could be performed on the languages that we defined.

We will begin our discussion of the context-free languages with another restricted formalism, the context-free grammar. But before we define it, we will pause and answer the more general question, "What is a grammar?"

11.1 Introduction to Rewrite Systems and Grammars

We'll begin with a very general computational model: Define a ***rewrite system*** (also called a ***production system*** or a ***rule-based system***) to be a list of rules and an algorithm for applying them. Each rule has a left-hand side and a right-hand side. For example, the following could be rewrite-system rules:

$$S \rightarrow aSb$$
$$aS \rightarrow \varepsilon$$
$$aSb \rightarrow bSabSa$$

In the discussion that follows, we will focus on rewrite system that operate on strings. But the core ideas that we will present can be used to define rewrite systems that operate on richer data structures. Of course, such data structures can be represented as strings, but the power of many practical rule-based systems comes from their ability to manipulate other structures directly.

Expert systems, (M.3.3) are programs that perform tasks in domains like engineering, medicine, and business, that require expertise when done by people. Many kinds of expertise can naturally be modeled as sets of condition/action rules. So many expert systems are built using tools that support rule-based programming.

Rule based systems are also used to model business practices (M.3.4) and as the basis for reasoning about the behavior of nonplayer characters in computer games. (N.3.3)

When a rewrite system R is invoked on some initial string w, it operates as follows:

simple-rewrite(R: rewrite system, w: initial string) =

1. Set *working-string* to w.
2. Until told by R to halt do:
   2.1. Match the left-hand side of some rule against some part of *working-string*.
   2.2. Replace the matched part of *working-string* with the right-hand side of the rule that was matched.
3. Return *working-string*.

If *simple-rewrite*(R, w) can return some string s then we'll say that R can ***derive*** s from w or that there exists a ***derivation*** in R of s from w.

Rewrite systems can model natural growth processes, as occur, for example, in plants. In addition, evolutionary algorithms can be applied to rule sets. Thus rewrite systems can model evolutionary processes. (Q.2.2)

We can define a particular ***rewrite-system formalism*** by specifying the form of the rules that are allowed and the algorithm by which they will be applied. In most of the rewrite-system formalisms that we will consider, a rule is simply a pair of strings. If the string on the left-hand side matches, it is replaced by the string on the right-hand side. But more flexible forms are also possible. For example, variables may be allowed. Let x be a variable. Then consider the rule:

$$axa \rightarrow aa$$

This rule will squeeze out whatever comes between a pair of a's.

Another useful form allows regular expressions as left-hand sides. If we do that, we can write rules like the following, which squeezes out b's between a's:

$$ab^*ab^*a \rightarrow aaa$$

The extended form of regular expressions that is supported in programming languages like Perl is often used to write substitution rules. (Appendix O)

In addition to describing the form of its rules, a rewrite-system formalism must describe how its rules will be applied. In particular, a rewrite-system formalism will define the conditions under which *simple-rewrite* will halt and the method by which it will choose a match in step 2.1. For example, one rewrite-system formalism might specify that any rule that matches may be chosen. A different formalism might specify that the rules have to be tried in the order in which they are written, with the first one that matches being the one that is chosen next.

Rewrite systems can be used to define functions. In this case, we write rules that operate on an input string to produce the required output string. Rewrite systems can also be used to define languages. In this case, we define a unique start symbol. The rules then apply and we will say that the language L that is generated by the system is exactly the set of strings, over L's alphabet, that can be derived by *simple-rewrite* from the start symbol.

> A rewrite-system formalism can be viewed as a programming language and some such languages turn out to be useful. For example, Prolog (M.2.3) supports a style of programming called logic programming. A logic program is a set of rules that correspond to logical statements of the form A if B. The interpreter for a logic program reasons backwards from a goal (such as A), chaining rules together until each right-hand side has been reduced to a set of facts (axioms) that are already known to be true.

The study of rewrite systems has played an important role in the development of the theory of computability. We'll see in Part V that there exist rewrite-system formalisms that have the same computational power as the Turing machine, both with respect to computing functions and with respect to defining languages. In the rest of our discussion in this chapter, however, we will focus just on their use to define languages.

A rewrite system that is used to define a language is called a ***grammar***. If G is a grammar, let $L(G)$ be the language that G generates. Like every rewrite system, every grammar contains a list (almost always treated as a set, i.e., as an unordered list) of rules. Also, like every rewrite system, every grammar works with an alphabet, which we can call V. In the case of grammars, we will divide V into two subsets:

- a ***terminal alphabet***, generally called Σ, which contains the symbols that make up the strings in $L(G)$, and
- a ***nonterminal alphabet***, the elements of which will function as working symbols that will be used while the grammar is operating. These symbols will disappear by the time the grammar finishes its job and generates a string.

One final thing is required to specify a grammar. Each grammar has a unique start symbol, often called S.

> Grammars can be used to describe phenomena as different as English (L.3), programming languages like Java (G.1), music (N.1), dance (Q.2.1), the growth of living organisms (Q.2.2), and the structure of RNA. (K.4)

A ***grammar formalism*** (like any rewrite-system formalism) specifies the form of the rules that are allowed and the algorithm by which they will be applied. The grammar formalisms that we will consider vary in the form of the rules that they allow. With one exception (Lindenmayer systems, which we'll describe in Section 24.4), all of the grammar formalisms that we will consider include a control algorithm that ignores rule order. Any rule that matches may be applied next.

To generate strings in $L(G)$, we invoke *simple-rewrite* (G, S). *Simple-rewrite* will begin with S and will apply the rules of G, which can be thought of (given the control algorithm we just described) as licenses to replace one string by another. At each step of one of its derivations, some rule whose left-hand side matches somewhere in *working-string* is selected. The substring that matched is replaced by the rule's right-hand side, generating a new value for *working string*.

> Grammars can be used to define languages that, in turn, define sets of things that don't look at all like strings. For example, SVG (Q.1.3) is a language that is used to describe two-dimensional graphics. SVG can be described with a context-free grammar.

We will use the symbol $\Rightarrow$ to indicate steps in a derivation. So, for example, suppose that G has the start symbol S and the rules $S \rightarrow$ aSb, $S \rightarrow$ bSa, and $S \rightarrow \varepsilon$. Then a derivation could begin with:

$$S \Rightarrow \text{a}S\text{b} \Rightarrow \text{aa}S\text{bb} \Rightarrow \ldots$$

At each step, it is possible that more than one rule's left-hand side matches the working string. It is also possible that a rule's left-hand side matches the working string in more than one way. In either case, there is a derivation corresponding to each alternative. It is precisely the existence of these choices that enables a grammar to generate more than one string.

Continuing with our example, there are three choices at the next step:

$S \Rightarrow \text{a}S\text{b} \Rightarrow \text{aa}S\text{bb} \Rightarrow \text{aaa}S\text{bbb}$ (using the first rule),
$S \Rightarrow \text{a}S\text{b} \Rightarrow \text{aa}S\text{bb} \Rightarrow \text{aab}S\text{abb}$ (using the second rule), and
$S \Rightarrow \text{a}S\text{b} \Rightarrow \text{aa}S\text{bb} \Rightarrow \text{aabb}$ (using the third rule).

The derivation process may end whenever one of the following things happens:

1. The working string no longer contains any nonterminal symbols (including, as a special case, when the working string is ε), or
2. There are nonterminal symbols in the working string but there is no match with the left-hand side of any rule in the grammar. For example, if the working string were $A\text{a}B\text{b}$, this would happen if the only left-hand side were C.

In the first case, but not the second, we say that the working string is ***generated*** by the grammar. Thus, the ***language*** that a grammar generates includes only strings over the terminal alphabet (i.e., strings in Σ^*). In the second case, we have a blocked or nonterminated derivation but no generated string.

It is also possible that, in a particular case, neither 1 nor 2 is achieved. Suppose, for example, that a grammar contained only the rules $S \rightarrow Ba$ and $B \rightarrow bB$, with S the start symbol. Then all derivations proceed in the following way:

$$S \Rightarrow Ba \Rightarrow bBa \Rightarrow bbBa \Rightarrow bbbBa \Rightarrow bbbbBa \Rightarrow \cdots$$

The working string is always rewriteable (in only one way, as it happens), and so this grammar can produce no terminated derivations consisting entirely of terminal symbols (i.e., generated strings). Thus this grammar generates the language $\emptyset$.

11.2 Context-Free Grammars and Languages

We've already seen our first specific grammar formalism. In Chapter 7, we defined a regular grammar to be one in which every rule must:

- have a left-hand side that is a single nonterminal, and
- have a right-hand side that is ε or a single terminal or a single terminal followed by a single nonterminal.

We now define a ***context-free grammar*** (or CFG) to be a grammar in which each rule must:

- have a left-hand side that is a single nonterminal, and
- have a right-hand side.

To simplify the discussion that follows, define an A rule, for any nonterminal symbol A, to be a rule whose left-hand side is A.

Next we must define a control algorithm of the sort we described at the end of the last section. A derivation will halt whenever no rule's left-hand side matches against *working-string*. At every step, any rule that matches may be chosen.

Context-free grammar rules may have any (possibly empty) sequence of symbols on the right-hand side. Because the rule format is more flexible than it is for regular grammars, the rules are more powerful. We will soon show some examples of languages that can be generated with context-free grammars but that can not be generated with regular ones.

All of the following are allowable context-free grammar rules (assuming appropriate alphabets):

$$S \rightarrow aSb$$
$$S \rightarrow \varepsilon$$
$$T \rightarrow T$$
$$S \rightarrow aSbbTT$$

The following are not allowable context-free grammar rules:

$$ST \rightarrow aSb$$
$$a \rightarrow aSb$$
$$\varepsilon \rightarrow a$$

The name for these grammars, "context-free," makes sense because, using these rules, the decision to replace a nonterminal by some other sequence is made without looking at the context in which the nonterminal occurs. In Chapters 23 and 24 we will consider less restrictive grammar formalisms in which the left-hand sides of the rules

may contain several symbols. For example, the rule $aSa \rightarrow aTa$ would be allowed. This rule says that S can be replaced by T when it is surrounded by a's. One of those formalisms is called "context-sensitive" because its rules allow context to be considered.

> Programming language syntax is typically described using context-free grammars, as we'll see below and in Appendix G.

Formally, a context-free grammar G is a quadruple (V, Σ, R, S), where:

- V is the rule alphabet, which contains nonterminals (symbols that are used in the grammar but that do not appear in strings in the language) and terminals,
- Σ (the set of terminals) is a subset of V,
- R (the set of rules) is a finite subset of $(V - \Sigma) \times V^*$, and
- S (the start symbol) can be any element of $V - \Sigma$.

Given a grammar G, define $x \Rightarrow_G y$ (abbreviated $\Rightarrow$ when G is clear from context) to be the binary relation *derives-in-one-step*, defined so that:

$$\forall x, y \in V^* (x \Rightarrow_G y \text{ iff } x = \alpha A \beta, y = \alpha \gamma \beta, \text{ and there exists a rule } A \rightarrow \gamma \text{ in } R_G).$$

Any sequence of the form $w_0 \Rightarrow_G w_1 \Rightarrow_G w_2 \Rightarrow_G \ldots \Rightarrow_G w_n$ is called a ***derivation*** in G. Let $\Rightarrow_G^*$ be the reflexive, transitive closure of $\Rightarrow_G$. We'll call $\Rightarrow_G^*$ the ***derives*** relation.

The ***language generated by G***, denoted $L(G)$, is $\{w \in \Sigma^* : S \Rightarrow_G^* w\}$. In other words, the language generated by G is the set of all strings of terminals that can be derived from S using zero or more applications of rules in G.

A language L is ***context-free*** iff it is generated by some context-free grammar G. The context-free languages (or CFLs) are a proper superset of the regular languages. In the next several examples, we will see languages that are context-free but not regular. Then, in Chapter 13, we will prove the other part of this claim, namely that every regular language is also context-free.

EXAMPLE 11.1 The Balanced Parentheses Language

Consider Bal = $\{w \in \{), (\}^* :$ the parentheses are balanced$\}$. We showed in Example 8.10 that Bal is not regular. But it is context-free because it can be generated by the grammar $G = \{\{S,), (\}, \{), (\}, R, S)$, where:

$$R = \{S \rightarrow (S)$$
$$S \rightarrow SS$$
$$S \rightarrow \varepsilon\}.$$

Some example derivations in G:

$$S \Rightarrow (S) \Rightarrow ().$$
$$S \Rightarrow (S) \Rightarrow (SS) \Rightarrow ((S)S) \Rightarrow (()S) \Rightarrow (()(S)) \Rightarrow (()()).$$

So, $S \Rightarrow^* ()$ and $S \Rightarrow^* (()())$.

The syntax of Boolean query languages is describable with a context-free grammar. (Q.11)

EXAMPLE 11.2 A^nB^n

Consider $A^nB^n = \{a^nb^n : n \geq 0\}$. We showed in Example 8.8 that A^nB^n is not regular. But it is context-free because it can be generated by the grammar $G = \{\{S, a, b\}, \{a, b\}, R, S\}$, where:

$$R = \{S \rightarrow aSb$$
$$S \rightarrow \varepsilon\}.$$

What is it about context-free grammars that gives them the power to define languages like Bal and A^nB^n?

We can begin answering that question by defining a rule in a grammar G to be ***recursive*** iff it is of the form $X \rightarrow w_1Yw_2$, where $Y \Rightarrow_G^* w_3Xw_4$ and all of w_1, w_2, w_3, and w_4 may be any element of V^*. A grammar is recursive iff it contains at least one recursive rule. For example, the grammar we just presented for *Bal* is recursive because it contains the rule $S \rightarrow (S)$. The grammar we presented for A^nB^n is recursive because it contains the rule $S \rightarrow aSb$. A grammar that contained the rule $S \rightarrow aS$ would also be recursive. So the regular grammar whose rules are $\{S \rightarrow aT, T \rightarrow aW, W \rightarrow aS, W \rightarrow a\}$ is recursive. Recursive rules make it possible for a finite grammar to generate an infinite set of strings.

Let's now look at an important property that gives context-free grammars the power to define languages that aren't regular. A rule in a grammar G is ***self-embedding*** iff it is of the form $X \rightarrow w_1Yw_2$, where $Y \Rightarrow_G^* w_3Xw_4$ and both w_1w_3 and w_4w_2 are in Σ^+. A grammar is self-embedding iff it contains at least one self-embedding rule. So now we require that a nonempty string be generated on each side of the nested X. The grammar we presented for *Bal* is self-embedding because it contains the rule $S \rightarrow (S)$. The grammar we presented for A^nB^n is self-embedding because it contains the rule $S \rightarrow aSb$. The presence of a rule like $S \rightarrow aS$ does not by itself make a grammar self-embedding. But the rule $S \rightarrow aT$ is self-embedding in any grammar G that also contains the rule $T \rightarrow Sb$, since $S \rightarrow aT$ and $T \Rightarrow_G^* Sb$. Self-embedding grammars are able to define languages like Bal, A^nB^n, and others whose strings must contain pairs of matching regions, often of the form uv^ixy^iz. No regular language can impose such a requirement on its strings.

The fact that a grammar G is self-embedding does not guarantee that $L(G)$ isn't regular. There might be a different grammar G' that also defines $L(G)$ and that is not self-embedding. For example, $G_1 = (\{S, a\}, \{a\}, \{S \rightarrow \varepsilon, S \rightarrow a, S \rightarrow aSa\}, S)$ is self-embedding, yet it defines the regular language a^*. However, we note the following two important facts:

- If a grammar G is not self-embedding then $L(G)$ is regular. Recall that our definition of regular grammars did not allow self-embedding.

- If a language L has the property that every grammar that defines it is self-embedding, then L is not regular.

The rest of the grammars that we will present in this chapter are self-embedding.

EXAMPLE 11.3 Even Length Palindromes

Consider PalEven $= \{ww^R : w \in \{a,b\}^*\}$, the language of even-length palindromes of a's and b's. We showed in Example 8.11 that PalEven is not regular. But it is context-free because it can be generated by the grammar $G = \{\{S, a, b\}, \{a, b\}, R, S\}$, where:

$$R = \{S \rightarrow aSa$$
$$S \rightarrow bSb$$
$$S \rightarrow \varepsilon\}.$$

EXAMPLE 11.4 Equal Numbers of a's and b's

Let $L = \{w \in \{a, b\}^* : \#_a(w) = \#_b(w)\}$. We showed in Example 8.14 that L is not regular. But it is context-free because it can be generated by the grammar $G = \{\{S, a, b\}, \{a, b\}, R, S\}$, where:

$$R = \{S \rightarrow aSb$$
$$S \rightarrow bSa$$
$$S \rightarrow SS$$
$$S \rightarrow \varepsilon\}.$$

These simple examples are interesting because they capture, in a couple of lines, the power of the context-free grammar formalism. But our real interest in context-free grammars comes from the fact that they can describe useful and powerful languages that are substantially more complex.

It quickly becomes apparent, when we start to build larger grammars, that we need a more flexible grammar-writing notation. We'll use the following two extensions when they are helpful:

- The symbol | should be read as "or". It allows two or more rules to be collapsed into one. So the following single rule is equivalent to the four rules we wrote in Example 11.4:

$$S \rightarrow aSb|bSa|SS|\varepsilon$$

- We often require nonterminal alphabets that contain more symbols than there are letters. To solve that problem, we will allow a nonterminal symbol to be any sequence of characters surrounded by angle brackets. So <program> and <variable> could be nonterminal symbols using this convention.

BNF (or Backus Naur form) is a widely used grammatical formalism that exploits both of these extensions. It was created in the late 1950s as a way to describe the programming language ALGOL 60. It has since been extended and several dialects developed. (G.1.1)

EXAMPLE 11.5 BNF for a Small Java Fragment

Because BNF was originally designed when only a small character set was available, it uses the three symbol sequence ::= in place of →. The following BNF-style grammar describes a highly simplified and very small subset of Java:

```
<block> ::= {<stmt-list>} | {}
<stmt-list> ::= <stmt> | <stmt-list> <stmt>
<stmt> ::= <block> | while (<cond>) <stmt> |
           if (<cond>) <stmt> |
           do <stmt> while (<cond>); | <assignment-stmt>; |
           return | return <expression> |
           <method-invocation>;
```

The rules of this grammar make it clear that the following block may be legal in Java (assuming that the appropriate declarations have occurred):

```
{     while (x < 12) {
            hippo.pretend(x);
            x = x + 2;
      }}
```

On the other hand, the following block is not legal:

```
{     while x < 12}) (
            hippo.pretend(x);
            x = x + 2;
      }}
```

Many other kinds of practical languages are also context-free. For example, HTML can be described with a context-free grammar using a BNF-style grammar. (Q.1.2)

EXAMPLE 11.6 A Fragment of an English Grammar

Much of the structure of an English sentence can be described by a (large) context-free grammar. For historical reasons, linguistic grammars typically use a

EXAMPLE 11.6 (*Continued*)

slightly different notational convention. Nonterminals will be written as strings whose first symbol is an upper case letter. So the following grammar describes a tiny fragment of English. The symbol *NP* will derive noun phrases; the symbol *VP* will derive verb phrases:

$S \rightarrow NP\ VP$
$NP \rightarrow$ the *Nominal* | a *Nominal* | *Nominal* | *ProperNoun* | *NP PP*
$Nominal \rightarrow N \mid Adjs\ N$
$N \rightarrow$ cat | dogs | bear | girl | chocolate | rifle
$ProperNoun \rightarrow$ Chris | Fluffy
$Adjs \rightarrow Adj\ Adjs \mid Adj$
$Adj \rightarrow$ young | older | smart
$VP \rightarrow V \mid V\ NP \mid VP\ PP$
$V \rightarrow$ like | likes | thinks | shot | smells
$PP \rightarrow Prep\ NP$
$Prep \rightarrow$ with

Is English (or German or Chinese) really context-free? (L.3.3)

11.3 Designing Context-Free Grammars

In this section, we offer a few simple strategies for designing straightforward context-free grammars. Later we'll see that some grammars are better than others (for various reasons) and we'll look at techniques for finding "good" grammars. For now, we will focus on finding some grammar.

The most important rule to remember in designing a context-free grammar to generate a language L is the following:

- If L has the property that every string in it has two regions and those regions must bear some relationship to each other (such as being of the same length), then the two regions must be generated in tandem. Otherwise, there is no way to enforce the necessary constraint.

Keeping that rule in mind, there are two simple ways to generate strings:

- To generate a string with multiple regions that must occur in some fixed order but do not have to correspond to each other, use a rule of the form:

$$A \rightarrow BC\ldots$$

This rule generates two regions, and the grammar that contains it will then rely on additional rules to describe how to form a B region and how to form a C region. Longer rules, like $A \rightarrow BCDE$, can be used if additional regions are necessary.

- To generate a string with two regions that must occur in some fixed order and that must correspond to each other, start at the outside edges of the string and generate toward the middle. If there is an unrelated region in between the related ones, it must be generated after the related regions have been produced.

> The outside-in structure of context-free grammars makes them well suited to describing physical things, like RNA molecules, that fold. (K.4)

EXAMPLE 11.7 Concatenating Independent Sublanguages

Let $L = \{a^n b^n c^m : n, m \geq 0\}$. Here, the c^m portion of any string in L is completely independent of the $a^n b^n$ portion, so we should generate the two portions separately and concatenate them together. So let $G = (\{S, N, C, a, b, c\}, \{a, b, c\}, R, S\}$ where:

$$
\begin{aligned}
R = \{&S \rightarrow NC && \text{/* Generate the two independent portions.}\\
&N \rightarrow aNb && \text{/* Generate the } a^n b^n \text{ portion, from the outside in.}\\
&N \rightarrow \varepsilon\\
&C \rightarrow cC && \text{/* Generate the } c^m \text{ portion.}\\
&C \rightarrow \varepsilon\}.
\end{aligned}
$$

EXAMPLE 11.8 The Kleene Star of a Language

Let $L = \{a^{n_1} b^{n_1} a^{n_2} b^{n_2} \ldots a^{n_k} b^{n_k} : k \geq 0 \text{ and } \forall i\ (n_i \geq 0)\}$. For example, the following strings are in L: ε, abab, aabbaaabbbabab. Note that $L = \{a^n b^n : n \geq 0\}^*$, which gives a clue how to write the grammar we need. We know how to produce individual elements of $\{a^n b^n : n \geq 0\}$, and we know how to concatenate regions together. So a solution is $G = (\{S, M, a, b\}, \{a, b\}, R, S\}$ where:

$$
\begin{aligned}
R = \{&S \rightarrow MS && \text{/* Each } M \text{ will generate one } \{a^n b^n : n \geq 0\}\\
& && \quad \text{region.}\\
&S \rightarrow \varepsilon\\
&M \rightarrow aMb && \text{/* Generate one region.}\\
&M \rightarrow \varepsilon\}.
\end{aligned}
$$

11.4 Simplifying Context-Free Grammars ●

In this section, we present two algorithms that may be useful for simplifying context-free grammars.

Consider the grammar $G = (\{S, A, B, C, D, a, b\}, \{a, b\}, R, S)$, where:

$$
\begin{aligned}
R = \{&S \rightarrow AB \mid AC\\
&A \rightarrow aAb \mid \varepsilon
\end{aligned}
$$

$B \rightarrow bA$
$C \rightarrow bCa$
$D \rightarrow AB\}$.

G contains two useless variables: C is useless because it is not able to generate any strings in Σ^*. (Every time a rule is applied to a C, a new C is added.) D is useless because it is unreachable, via any derivation, from S. So any rules that mention either C or D can be removed from G without changing the language that is generated. We present two algorithms, one to find and remove variables like C that are unproductive, and one to find and remove variables like D that are unreachable.

Given a grammar $G = (V, \Sigma, R, S)$, we define *removeunproductive*(G) to create a new grammar G', where $L(G') = L(G)$ and G' does not contain any unproductive symbols. Rather than trying to find the unproductive symbols directly, *removeunproductive* will find and mark all the productive ones. Any that are left unmarked at the end are unproductive. Initially, all terminal symbols will be marked as productive since each of them generates a terminal string (itself). A nonterminal symbol will be marked as productive when it is discovered that there is at least one way to rewrite it as a sequence of productive symbols. So *removeunproductive* effectively moves backwards from terminals, marking nonterminals along the way.

removeunproductive$(G$: CFG) =

1. $G' = G$.
2. Mark every nonterminal symbol in G' as unproductive.
3. Mark every terminal symbol in G' as productive.
4. Until one entire pass has been made without any new symbol being marked do:

   For each rule $X \rightarrow \alpha$ in R do:

   If every symbol in α has been marked as productive and X has not yet been marked as productive, then mark X as productive.
5. Remove from $V_{G'}$ every unproductive symbol.
6. Remove from $R_{G'}$ every rule with an unproductive symbol on either the left-hand side or the right-hand side.
7. Return G'

Removeunproductive must halt because there is only some finite number of nonterminals that can be marked as productive. So the maximum number of times it can execute step 4 is $|V - \Sigma|$. Clearly $L(G') \subseteq L(G)$ since G' can produce no derivations that G could not have produced. And $L(G') = L(G)$ because the only derivations that G can perform but G' cannot are those that do not end with a terminal string.

Notice that it is possible that S is unproductive. This will happen precisely in case $L(G) = \varnothing$. We will use this fact in Section 14.1.2 to show the existence of a procedure that decides whether or not a context-free language is empty.

Next we'll define an algorithm for getting rid of unreachable symbols like D in the grammar we presented above. Given a grammar $G = (V, \Sigma, R, S)$, we define *removeunreachable*(G) to create a new grammar G', where $L(G') = L(G)$ and G'

does not contain any unreachable nonterminal symbols. What *removeunreachable* does is to move forward from S, marking reachable symbols along the way.

removeunreachable(G: CFG) =

1. $G' = G$.
2. Mark S as reachable.
3. Mark every other nonterminal symbol as unreachable.
4. Until one entire pass has been made without any new symbol being marked do:
   For each rule $X \rightarrow \alpha A \beta$ (where $A \in V - \Sigma$ and $\alpha, \beta \in V^*$) in R do:
   If X has been marked as reachable and A has not, then mark A as reachable.
5. Remove from $V_{G'}$ every unreachable symbol.
6. Remove from $R_{G'}$ every rule with an unreachable symbol on the left-hand side.
7. Return G'.

Removeunreachable must halt because there is only some finite number of nonterminals that can be marked as reachable. So the maximum number of times it can execute step 4 is $|V - \Sigma|$. Clearly $L(G') \subseteq L(G)$ since G' can produce no derivations that G could not have produced. And $L(G') = L(G)$ because every derivation that can be produced by G can also be produced by G'.

11.5 Proving That a Grammar is Correct ●

In the last couple of sections, we described some techniques that are useful in designing context-free languages and we argued that the grammars that we built were correct (i.e., that they correctly describe languages with certain properties). But, given some language L and a grammar G, can we actually prove that G is correct (i.e., that it generates exactly the strings in L)? To do so, we need to prove two things:

1. G generates only strings in L, and
2. G generates all the strings in L.

The most straightforward way to do step 1 is to imagine the process by which G generates a string as the following loop (a version of *simple-rewrite*, using *st* in place of *working-string*):

1. $st = S$.
2. Until no nonterminals are left in *st* do:
   Apply some rule in R to *st*.
3. Output *st*.

Then we construct a loop invariant I and show that:

- I is true when the loop begins,
- I is maintained at each step through the loop (i.e., by each rule application), and
- $I \wedge$ (*st* contains only terminal symbols) $\rightarrow st \in L$.

Step 2 is generally done by induction on the length of the generated strings.

EXAMPLE 11.9 The Correctness of the A^nB^n Grammar

In Example 11.2, we considered the language A^nB^n. We built for it the grammar $G = \{\{S, a, b\}, \{a, b\}, R, S\}$, where:

$$R = \{S \rightarrow aSb \quad (1)$$
$$S \rightarrow \varepsilon\}. \quad (2)$$

We now show that G is correct. We first show that every string w in $L(G)$ is in A^nB^n: Let st be the working string at any point in a derivation in G. We need to define I so that it captures the two features of every string in A^nB^n: The number of a's equals the number of b's and the letters are in the correct order. So we let I be:

$$(\#_a(st) = \#_b(st)) \wedge (st \in a^*(S \cup \varepsilon)b^*).$$

Now we prove:

- I is true when $st = S$: In this case, $\#_a(st) = \#_b(st)) = 0$ and st is of the correct form.
- If I is true before a rule fires, then it is true after the rule fires: To prove this, we consider the rules one at a time and show that each of them preserves I. Rule (1) adds one a and one b to st, so it does not change the difference between the number of a's and the number of b's. Further, it adds the a to the left of S and the b to the right of S, so if the form constraint was satisfied before applying the rule it still is afterwards. Rule (2) adds nothing so it does not change either the number of a's or b's or their locations.
- If I is true and st contains only terminal symbols, then $st \in A^nB^n$: In this case, st possesses the three properties required of all strings in A^nB^n: They are composed only of a's and b's, $(\#_a(st) = \#_b(st))$, and all a's come before all b's.

Next we show that every string w in A^nB^n can be generated by G: Every string in A^nB^n is of even length, so we will prove the claim only for strings of even length. The proof is by induction on $|w|$:

- Base case: If $|w| = 0$, then $w = \varepsilon$, which can be generated by applying rule (2) to S.
- Prove: If every string in A^nB^n of length k, where k is even, can be generated by G, then every string in A^nB^n of length $k + 2$ can also be generated. Notice that, for any even k, there is exactly one string in A^nB^n of length k : $a^{k/2}b^{k/2}$. There is also only one string of length $k + 2$, namely $aa^{k/2}b^{k/2}b$, that can be generated by first applying rule (1) to produce aSb, and then applying to S whatever rule sequence generated $a^{k/2}b^{k/2}$. By the induction hypothesis, such a sequence must exist.

EXAMPLE 11.10 The Correctness of the Equal a's and b's Grammar

In Example 11.4 we considered the language $L = \{w \in \{a, b\}^* : \#_a(w) = \#_b(w)\}$. We built for it the grammar $G = \{\{S, a, b\}, \{a, b\}, R, S\}$, where:

$$\begin{aligned} R = \{ & S \rightarrow aSb \quad (1) \\ & S \rightarrow bSa \quad (2) \\ & S \rightarrow SS \quad (3) \\ & S \rightarrow \varepsilon\}. \quad (4) \end{aligned}$$

This time it is perhaps less obvious that G is correct. In particular, does it generate every sequence where the number of a's equals the number of b's? The answer is yes, which we now prove.

To make it easy to describe this proof, we define the following function:

$$\Delta(w) = \#_a(w) - \#_b(w).$$

Note that a string w is in L iff $w \in \{a, b\}^*$ and $\Delta(w) = 0$.

We begin by showing that every string w in $L(G)$ is in L: Again, let st be the working string at any point in a derivation in G. Let I be:

$$st \in \{a, b, S\}^* \wedge \Delta(st) = 0.$$

Now we prove:

- I is true when $st = S$: In this case, $\#_a(st) = \#_b(st)) = 0$. So $\Delta(st) = 0$.
- If I is true before a rule fires, then it is true after the rule fires: The only symbols that can be added by any rule are a, b, and S. Rules (1) and (2) each add one a and one b to st, so neither of them changes $\Delta(st)$. Rules (3) and (4) add neither a's nor b's to the working string, so $\Delta(st)$ does not change.
- If I is true and st contains only terminal symbols, then $st \in L$: In this case, st possesses the two properties required of all strings in L: They are composed only of a's and b's and $\Delta(st) = 0$.

It is perhaps less obviously true that G generates every string in L. Can we be sure that there are no permutations that it misses? Yes, we can. We next we show that every string w in L can be generated by G. Every string in L is of even length, so we will prove the claim only for strings of even length. The proof is by induction on $|w|$.

- Base case: If $|w| = 0$, $w = \varepsilon$, which can be generated by applying rule (4) to S.
- Prove that if every string in L of length $\leq k$, where k is even, can be generated by G, then every string w in L of length $k + 2$ can also be generated: Since w has length $k + 2$, it can be rewritten as one of the following: axb, bxa, axa, or bxb, for some $x \in \{a, b\}^*$. $|x| = k$. We consider two cases:
  - w = axb or bxa. If $w \in L$, then $\Delta(w) = 0$ and so $\Delta(x)$ must also be 0. $|x| = k$. So, by the induction hypothesis, G generates x. Thus G can also generate w: It first applies either rule (1) (if w = axb) or rule (2) (if w = bxa). It then applies to S whatever rule sequence generated x. By the induction hypothesis, such a sequence must exist.

EXAMPLE 11.10 (*Continued*)

- w = axa, or bxb. We consider the former case. The argument is parallel for the latter. Note that any string in L, of either of these forms, must have length at least 4. We will show that $w = vy$, where both v and y are in L, $2 \leq |v| \leq k$, and $2 \leq |y| \leq k$. If that is so, then G can generate w by first applying rule (3) to produce SS, and then generating v from the first S and y from the second S. By the induction hypothesis, it must be possible for it to do that since both v and y have length $\leq k$.

  To find v and y, we can imagine building w (which we've rewritten as axa) up by concatenating one character at a time on the right. After adding only one character, we have just a. $\Delta(\text{a}) = 1$. Since $w \in L$, $\Delta(w) = 0$. So $\Delta(\text{ax}) = -1$ (since it is missing the final a of w). The value of Δ changes by exactly 1 each time a symbol is added to a string. Since Δ is positive when only a single character has been added and becomes negative by the time the string ax has been built, it must at some point before then have been 0. Let v be the shortest nonempty prefix of w to have a value of 0 for Δ. Since v is nonempty and only even length strings can have Δ equal to 0, $2 \leq |v|$. Since Δ became 0 sometime before w became ax, v must be at least two characters shorter than w (it must be missing at least the last character of x plus the final a), so $|v| \leq k$. Since $\Delta(v) = 0$, $v \in L$. Since $w = vy$, we know bounds on the length of y: $2 \leq |y| \leq k$. Since $\Delta(w) = 0$ and $\Delta(v) = 0$, $\Delta(y)$ must also be 0 and so $y \in L$.

11.6 Derivations and Parse Trees

Context-free grammars do more than just describe the set of strings in a language. They provide a way of assigning an internal structure to the strings that they derive. This structure is important because it, in turn, provides the starting point for assigning meanings to the strings that the grammar can produce.

The grammatical structure of a string is captured by a ***parse tree***, which records which rules were applied to which nonterminals during the string's derivation. In Chapter 15, we will explore the design of programs, called ***parsers***, that, given a grammar G and a string w, decide whether $w \in L(G)$ and, if it is, create a parse tree that captures the process by which G could have derived w.

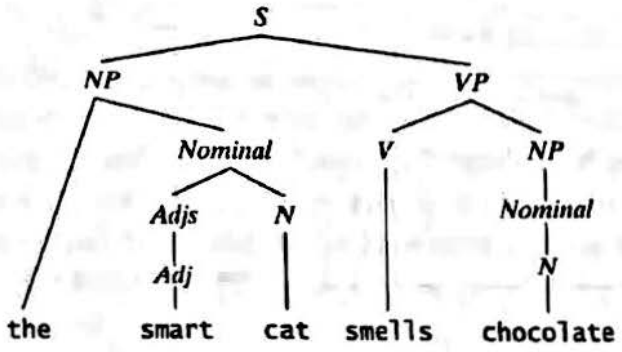
A parse tree, derived by a grammar $G = (V, \Sigma, R, S)$, is a rooted, ordered tree in which:

- Every leaf node is labeled with an element of $\Sigma \cup \{\varepsilon\}$,
- The root node is labeled S,
- Every other node is labeled with some element of $V - \Sigma$, and
- If m is a nonleaf node labeled X and the children of m are labeled $x_1, x_2, \ldots, x_n$, then R contains the rule $X \rightarrow x_1, x_2, \ldots, x_n$.

Define the ***branching factor*** of a grammar G to be length (the number of symbols) of the longest right-hand side of any rule in G. Then the branching factor of any parse tree generated by G is less than or equal to the branching factor of G.

EXAMPLE 11.11 The Parse Tree of a Simple English Sentence

Consider again the fragment of an English grammar that we wrote in Example 11.6. That grammar can be used to produce the following parse tree for the sentence `the smart cat smells chocolate`:



Notice that, in Example 11.11, the constituents (the subtrees) correspond to objects (like some particular cat) that have meaning in the world that is being described. It is clear from the tree that this sentence is not about cat smells or smart cat smells.

Because parse trees matter, it makes sense, given a grammar G, to distinguish between:

- G's ***weak generative capacity***, defined to be the set of strings, $L(G)$, that G generates, and
- G's ***strong generative capacity***, defined to be the set of parse trees that G generates.

When we design grammars it will be important that we consider both their weak and their strong generative capacities.

In our last example, the process of deriving the sentence `the smart cat smells chocolate` began with:

$$S \Rightarrow NP\ VP \Rightarrow \ldots$$

Looking at the parse tree, it isn't possible to tell which of the following happened next:

$$S \Rightarrow NP\ VP \Rightarrow \text{The } Nominal\ VP \Rightarrow$$
$$S \Rightarrow NP\ VP \Rightarrow NP\ V\ NP \Rightarrow$$

Parse trees are useful precisely because they capture the important structural facts about a derivation but throw away the details of the order in which the nonterminals were expanded.

While it's true that the order in which nonterminals are expanded has no bearing on the structure that we wish to assign to a string, order will become important when

we attempt to define algorithms that work with context-free grammars. For example, in Chapter 15 we will consider various parsing algorithms for context-free languages. Given an input string w, such algorithms must work systematically through the space of possible derivations in search of one that could have generated w. To make it easier to describe such algorithms, we will define two useful families of derivations:

- A ***left-most derivation*** is one in which, at each step, the leftmost nonterminal in the working string is chosen for expansion.
- A ***right-most derivation*** is one in which, at each step, the rightmost nonterminal in the working string is chosen for expansion.

Returning to the `smart cat` example above:

- A left-most derivation is:

$S \Rightarrow NP\ VP \Rightarrow$ The $Nominal\ VP \Rightarrow$ The $Adjs\ N\ VP \Rightarrow$ The $Adj\ N\ VP \Rightarrow$
The smart $N\ VP \Rightarrow$ the smart cat $VP \Rightarrow$ the smart cat $V\ NP \Rightarrow$
the smart cat smells $NP \Rightarrow$ the smart cat smells $Nominal \Rightarrow$
the smart cat smells $N \Rightarrow$ the smart cat smells chocolate

- A right-most derivation is:

$S \Rightarrow NP\ VP \Rightarrow NP\ V\ NP \Rightarrow NP\ V\ Nominal \Rightarrow NP\ V\ N \Rightarrow NP\ V$ chocolate $\Rightarrow$
NP smells chocolate $\Rightarrow$ the $Nominal$ smells chocolate $\Rightarrow$
the $Adjs\ N$ smells chocolate $\Rightarrow$ The $Adjs$ cat smells chocolate $\Rightarrow$
the Adj cat smells chocolate $\Rightarrow$ the smart cat smells chocolate

11.7 Ambiguity

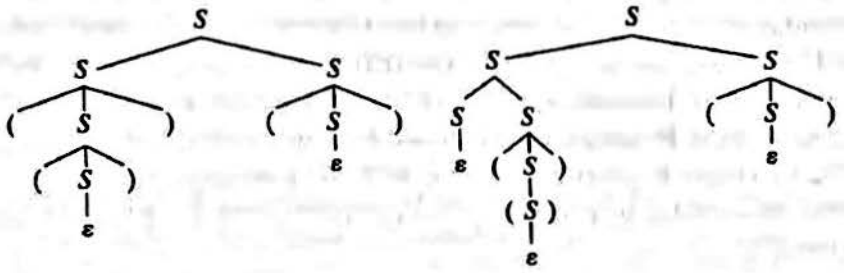
Sometimes a grammar may produce more than one parse tree for some (or all) of the strings it generates. When this happens, we say that the grammar is ambiguous. More precisely, a grammar G is ***ambiguous*** iff there is at least one string in $L(G)$ for which G produces more than one parse tree. It is easy to write ambiguous grammars if we are not careful. In fact, we already have.

EXAMPLE 11.12 The Balanced Parentheses Grammar is Ambiguous

Recall the language Bal = $\{w \in \{), (\}^* :$ the parentheses are balanced$\}$, for which we wrote the grammar $G = (\{S,), (\}, \{), (\}, R, S)$, where:

$$R = \{S \rightarrow (S)$$
$$S \rightarrow SS$$
$$S \rightarrow \varepsilon\}.$$

G can produce both of the following parse trees for the string (())():



In fact, G can produce an infinite number of parse trees for the string (())().

A grammar G is unambiguous iff, for all strings w, at every point in a leftmost or rightmost derivation of w, only one rule in G can be applied. The grammar that we just presented in Example 11.12 clearly fails to meet this requirement. For example, here are two leftmost derivations of the string (())():

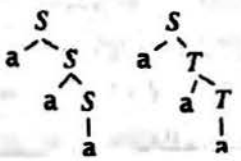
- $S \Rightarrow SS \Rightarrow (S)S \Rightarrow ((S))S \Rightarrow (())S \Rightarrow (())(S) \Rightarrow (())()$.
- $S \Rightarrow SS \Rightarrow SSS \Rightarrow SS \Rightarrow (S)S \Rightarrow ((S))S \Rightarrow (())S \Rightarrow (())(S) \Rightarrow (())()$.

11.7.1 Why Is Ambiguity a Problem?

Why are we suddenly concerned with ambiguity? Regular grammars can also be ambiguous. And regular expressions can often derive a single string in several distinct ways.

EXAMPLE 11.13 Regular Expressions and Grammars Can Be Ambiguous

Let $L = \{w \in \{a, b\}^* : w \text{ contains at least one a}\}$. L is regular. It can be defined with both a regular expression and a regular grammar. We show two ways in which the string aaa can be generated from the regular expression we have written and two ways in which it can be generated by the regular grammar:

Regular Expression	**Regular Grammar**
(a ∪ b)*a (a∪ b)*.	$S \rightarrow a$
	$S \rightarrow bS$
choose a from (a∪ b), then	$S \rightarrow aS$
choose a from (a∪ b), then	$S \rightarrow aT$
choose a, then	$T \rightarrow a$
choose ε from (a∪ b)*.	$T \rightarrow b$
	$T \rightarrow aT$
or	$T \rightarrow bT$
choose ε from (a∪ b)*, then	
choose a, then	
choose a from (a∪ b), then	
choose a from (a∪ b).	

We had no reason to be concerned with ambiguity when we were discussing regular languages because, for most applications of them, we don't care about assigning internal structure to strings. With context-free languages, we usually do care about internal structure because, given a string w, we want to assign meaning to w. We almost always want to assign a unique such meaning. It is generally difficult, if not impossible, to assign a unique meaning without a unique parse tree. So an ambiguous grammar, which fails to produce a unique parse tree, is a problem, as we'll see in our next example.

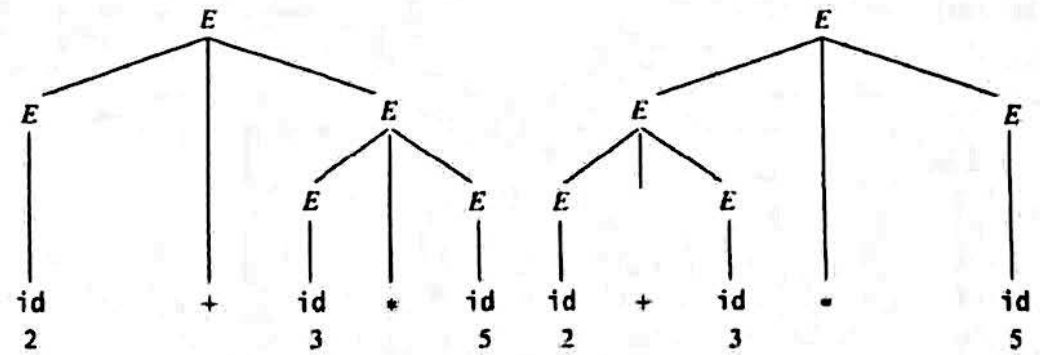
EXAMPLE 11.14 An Ambiguous Expression Grammar

Consider E_{xpr}, which we'll define to be the language of simple arithmetic expressions of the kind that could be part of anything from a small calculator to a programming language. We can define E_{xpr} with the following context-free grammar $G = \{\{E, \texttt{id}, +, *, (,)\}, \{\texttt{id}, +, *, (,)\}, R, E\}$, where:

$$R = \{E \rightarrow E + E$$
$$E \rightarrow E * E$$
$$E \rightarrow (E)$$
$$E \rightarrow \texttt{id}\}.$$

So that we can focus on the issues we care about, we've used the terminal symbol id as a shorthand for any of the numbers or variables that can actually occur as the operands in the expressions that G generates. Most compilers and interpreters for expression languages handle the parsing of individual operands in a first pass, called lexical analysis, which can be done with an FSM. We'll return to this topic in Chapter 15.

Consider the string 2 + 3 * 5, which we will write as id + id * id. Using G, we can get two parses for this string:



Should an evaluation of this expression return 17 or 25? (See Example 11.19 for a different expression grammar that fixes this problem.)

> Natural languages, like English and Chinese, are not explicitly designed. So it isn't possible to go in and remove ambiguity from them. See Example 11.22 and L.3.4.

Designers of practical languages must be careful that they create languages for which they can write unambiguous grammars.

11.7.2 Inherent Ambiguity

In many cases, when confronted with an ambiguous grammar G, it is possible to construct a new grammar G' that generates $L(G)$ and that has less (or no) ambiguity. Unfortunately, it is not always possible to do this. There exist context-free languages for which no unambiguous grammar exists. We call such languages ***inherently ambiguous***.

EXAMPLE 11.15 An Inherently Ambiguous Language

Let $L = \{a^i b^j c^k : i, j, k \geq 0, i = j \text{ or } j = k\}$. An alternative way to describe it is $\{a^n b^n c^m : n, m \geq 0\} \cup \{a^n b^m c^m : n, m \geq 0\}$. Every string in L has either (or both) the same number of a's and b's or the same number of b's and c's. L is inherently ambiguous. One grammar that describes it is $G = (\{S, S_1, S_2, A, B, a, b, c\}, \{a, b, c\}, R, S\}$, where:

$$
\begin{aligned}
R = \{ & S \rightarrow S_1 \mid S_2 \\
& S_1 \rightarrow S_1 c \mid A && \text{/* Generate all strings in } \{a^n b^n c^m : n, m \geq 0\}. \\
& A \rightarrow aAb \mid \varepsilon \\
& S_2 \rightarrow aS_2 \mid B && \text{/* Generate all strings in } \{a^n b^m c^m : n, m \geq 0\}. \\
& B \rightarrow bBc \mid \varepsilon\}.
\end{aligned}
$$

Now consider the strings in $A^n B^n C^n = \{a^n b^n c^n : n \geq 0\}$. They have two distinct derivations, one through S_1 and the other through S_2. It is possible to prove that L is inherently ambiguous: Given any grammar G that generates L there is at least one string with two derivations in G.

EXAMPLE 11.16 Another Inherently Ambiguous Language

Let $L = \{a^i b^j a^k b^l : i, j, k, l \geq 0, i = k \text{ or } j = l\}$. L is also inherently ambiguous.

Unfortunately, there are no clean fixes for the ambiguity problem for context-free languages. In Section 22.5 we'll see that both of the following problems are undecidable:

- Given a context-free grammar G, is G ambiguous?
- Given a context-free language L, is L inherently ambiguous?

11.7.3 Techniques for Reducing Ambiguity ♠

Despite the negative theoretical results that we have just mentioned, it is usually very important, when we are designing practical languages and their grammars, that we come up with a language that is not inherently ambiguous and a grammar for it that is unambiguous. Although there exists no general purpose algorithm to test for ambiguity in a grammar or to remove it when it is found (since removal is not always possible), there do exist heuristics that we can use to find some of the more common sources of ambiguity and remove them. We'll consider here three grammar structures that often lead to ambiguity:

1. ε rules like $S \rightarrow \varepsilon$.
2. Rules like $S \rightarrow SS$ or $E \rightarrow E + E$. In other words recursive rules whose right-hand sides are symmetric and contain at least two copies of the nonterminal on the left-hand side.
3. Rule sets that lead to ambiguous attachment of optional postfixes.

Eliminating ε-Rules

In Example 11.12, we showed a grammar for the balanced parentheses language. That grammar is highly ambiguous. Its major problem is that it is possible to apply the rule $S \rightarrow SS$ arbitrarily often, generating unnecessary instances of S, which can then be wiped out without a trace using the rule $S \rightarrow \varepsilon$. If we could eliminate the rule $S \rightarrow \varepsilon$, we could eliminate that source of ambiguity. We'll call any rule whose right-hand side is ε an ε-rule.

We'd like to define an algorithm that could remove ε-rules from a grammar G without changing the language that G generates. Clearly if $\varepsilon \in L(G)$, that won't be possible. Only an ε-rule can generate ε. However, it is possible to define an algorithm that eliminates ε-rules from G and leaves $L(G)$ unchanged except that, if $\varepsilon \in L(G)$, it will be absent from the language generated by the new grammar. We will show such an algorithm. Then we'll show a simple way to add ε back in, when necessary, without adding back the kind of ε-rules that cause ambiguity.

Let $G = (V, \Sigma, R, S)$ be any context-free grammar. The following algorithm constructs a new grammar G' such that $L(G') = L(G) - \{\varepsilon\}$ and G' contains no ε-rules:

removeEps (G: CFG) =

1. Let $G' = G$.
2. Find the set N of nullable variables in G'. A variable X is ***nullable*** iff either:
   (1) there is a rule $X \rightarrow \varepsilon$, or
   (2) there is a rule $X \rightarrow PQR\ldots$ such that $P, Q, R, \ldots$ are all nullable.

   So compute N as follows:
   2.1. Set N to the set of variables that satisfy (1).
   2.2. Until an entire pass is made without adding anything to N do:

   Evaluate all other variables with respect to (2). If any variable satisfies (2) and is not in N, insert it.

3. Define a rule to be ***modifiable*** iff it is of the form $P \rightarrow \alpha Q \beta$ for some Q in N and any α, β in V^*. Since Q is nullable, it could be wiped out by the application of ε-rules. But those rules are about to be deleted. So one possibility should be that Q just doesn't get generated in the first place. To make that happen requires adding new rules. So, repeat until G' contains no modifiable rules that haven't been processed:
   3.1. Given the rule $P \rightarrow \alpha Q \beta$, where $Q \in N$, add the rule $P \rightarrow \alpha\beta$ if it is not already present and if $\alpha\beta \neq \varepsilon$ and if $P \neq \alpha\beta$. This last check prevents adding the useless rule $P \rightarrow P$, which would otherwise be generated if the original grammar contained, for example, the rule $P \rightarrow PQ$ and Q were nullable.
4. Delete from G' all rules of the form $X \rightarrow \varepsilon$.
5. Return G'.

If *removeEps* halts, $L(G') = L(G) - \{\varepsilon\}$ and G' contains no ε-rules. And *removeEps* must halt. Since step 2 must add a nonterminal to N at each pass and it cannot add any symbol more than once, it must halt within $|V - \Sigma|$ passes. Step 3 may have to be done once for every rule in G and once for every new rule that it adds. But note that, whenever it adds a new rule, that rule has a shorter right-hand side than the rule from which it came. So the number of new rules that can be generated by some original rule in G is finite. So step 3 can execute only a finite number of times.

EXAMPLE 11.17 Eliminating ε-Rules

Let $G = (\{S, T, A, B, C, \mathrm{a}, \mathrm{b}, \mathrm{c}\}, \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}, R, S)$, where:

$$
\begin{aligned}
R = \{ & S \rightarrow \mathrm{a}T\mathrm{a} \\
& T \rightarrow ABC \\
& A \rightarrow \mathrm{a}A \mid C \\
& B \rightarrow B\mathrm{b} \mid C \\
& C \rightarrow \mathrm{c} \mid \varepsilon \}.
\end{aligned}
$$

On input G, *removeEps* behaves as follows: Step 2 finds the set N of nullable variables by initially setting N to $\{C\}$. On its first pass through step 2.2 it adds A and B to N. On the next pass, it adds T (since now A, B, and C are all in N). On the next pass, no new elements are found, so step 2 halts with $N = \{C, A, B, T\}$. Step 3 adds the following new rules to G':

$S \rightarrow \mathrm{aa}$	/* Since T is nullable.
$T \rightarrow BC$	/* Since A is nullable.
$T \rightarrow AC$	/* Since B is nullable.
$T \rightarrow AB$	/* Since C is nullable.
$T \rightarrow C$	/* From $T \rightarrow BC$, since B is nullable. Or from $T \rightarrow AC$.
$T \rightarrow B$	/* From $T \rightarrow BC$, since C is nullable. Or from $T \rightarrow AB$.

EXAMPLE 11.17 (*Continued*)

$T \rightarrow A$ /* From $T \rightarrow AC$, since C is nullable. Or from $T \rightarrow AB$.
$A \rightarrow$ a /* Since A is nullable.
$B \rightarrow$ b /* Since B is nullable.

Finally, step 4 deletes the rule $C \rightarrow \varepsilon$.

Sometimes $L(G)$ contains ε and it is important to retain it. To handle this case, we present the following algorithm, which constructs a new grammar G'', such that $L(G'') = L(G)$. If $L(G)$ contains ε, then G'' will contain a single ε-rule that can be thought of as being "quarantined". Its sole job is to generate the string ε. It can have no interaction with the other rules of the grammar.

atmostoneEps (G: CFG) =

1. $G'' = removeEps(G)$.
2. If S_G is nullable then: /* This means that $\varepsilon \in L(G)$.
   2.1. Create in G'' a new start symbol S^*.
   2.2. Add to $R_{G'}$ the two rules: $S^* \rightarrow \varepsilon$ and $S^* \rightarrow S_G$.
3. Return G''.

EXAMPLE 11.18 Eliminating ε-Rules from the Balanced Parens Grammar

We again consider Bal = $\{w \in \{), (\}^* :$ the parentheses are balanced$\}$ and the grammar $G = \{\{S,), (\}, \{), (\}, R, S)$, where:

$$R = \{S \rightarrow (S) \quad (1)$$
$$S \rightarrow SS \quad (2)$$
$$S \rightarrow \varepsilon\}. \quad (3)$$

We would like to eliminate the ambiguity in G. Since $\varepsilon \in L(G)$, we call *atmostoneEps*(G), which begins by applying *removeEps* to G:

- In step 2, $N = \{S\}$.
- In step 3, rule (1) causes us to add the rule $S \rightarrow ()$. Rule (2) causes us to consider adding the rule $S \rightarrow S$, but we omit adding rules whose right-hand sides and left-hand sides are the same.
- In step 4, we delete the rule $S \rightarrow \varepsilon$.

So *removeEps*(G) returns the grammar $G' = \{\{S,), (\}, \{), (\}, R, S)$, where $R =$

$\{S \rightarrow (S)$
$S \rightarrow ()$
$S \rightarrow SS\}$.

In its step 2, *atmostoneEps* creates the new start symbol S^*. In step 3, it adds the two rules $S^* \rightarrow \varepsilon$, $S^* \rightarrow S$. So *atmostoneEps* returns the grammar $G'' = \{\{S^*, S,), (\}, \{), (\}, R, S^*)$, where:

$$\begin{aligned} R = \{ & S^* \rightarrow \varepsilon \\ & S^* \rightarrow S \\ & S \rightarrow (S) \\ & S \rightarrow () \\ & S \rightarrow SS\}. \end{aligned}$$

The string (())() has only one parse in G''.

Eliminating Symmetric Recursive Rules

The new grammar that we just built for Bal is better than our original one. But it is still ambiguous. The string ()()() has two parses, shown in Figure 11.1. The problem now is the rule $S \rightarrow SS$, which must be applied $n - 1$ times to generate a sequence of n balanced parentheses substrings. But, at each time after the first, there is a choice of which existing S to split.

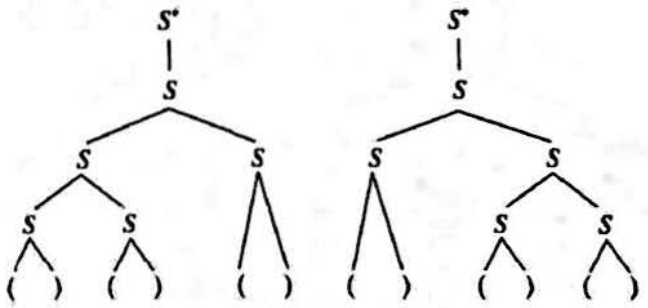


FIGURE 11.1 Two parse trees for the string ()()().

The solution to this problem is to rewrite the grammar so that there is no longer a choice. We replace the rule $S \rightarrow SS$ with one of the following rules:

$S \rightarrow SS_1$ /* force branching to the left.

$S \rightarrow S_1S$ /* force branching to the right.

Then we add the rule $S \rightarrow S_1$ and replace the rules $S \rightarrow (S)$ and $S \rightarrow ()$ with the rules $S_1 \rightarrow (S)$ and $S_1 \rightarrow ()$. What we have done is to change the grammar so that branching can occur only in one direction. Every S that is generated can branch, but no S_1 can. When all the branching has happened, S rewrites to S_1 and the rest of the derivation can occur.

So one unambiguous grammar for Bal is $G = \{\{S,), (\}, \{), (\}, R, S)$, where:

$$\begin{aligned} R = \{ & S^* \rightarrow \varepsilon && (1) \\ & S^* \rightarrow S && (2) \\ & S \rightarrow SS_1 && (3) \quad \text{/* Force branching to the left.} \\ & S \rightarrow S_1 && (4) \\ & S_1 \rightarrow (S) && (5) \\ & S_1 \rightarrow ()\}. && (6) \end{aligned}$$

The technique that we just used for Bal is useful in any situation in which ambiguity arises from a recursive rule whose right-hand side contains two or more copies of the left-hand side. An important application of this idea is to expression languages, like the language of arithmetic expressions that we introduced in Example 11.14.

EXAMPLE 11.19 An Unambiguous Expression Grammar

Consider again the language E_{xpr}, which we defined with the following context-free grammar $G = \{\{E, \texttt{id}, +, *, (,)\}, \{\texttt{id}, +, *, (,)\}, R, E\}$, where:

$$\begin{aligned} R = \{ & E \rightarrow E + E \\ & E \rightarrow E * E \\ & E \rightarrow (E) \\ & E \rightarrow \texttt{id} \}. \end{aligned}$$

G is ambiguous in two ways:

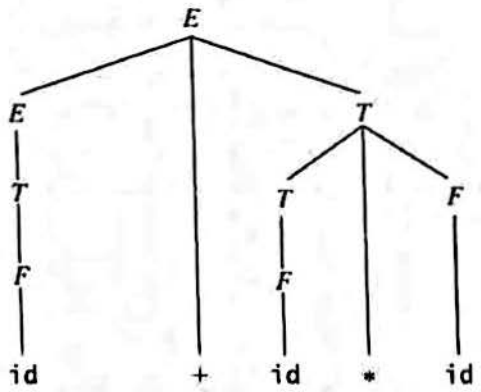
1. It fails to specify associativity. So, for example, there are two parses for the string `id + id + id`, corresponding to the bracketings `(id + id) + id` and `id + (id + id)`.
2. It fails to define a precedence hierarchy for the operators + and *. So, for example, there are two parses for the string `id + id * id`, corresponding to the bracketings `(id + id) * id` and `id + (id * id)`.

The first of these problems is analogous to the one we just solved for Bal. We could apply that solution here, but then we'd still have the second problem. We can solve both of them with the following grammar $G' = \{\{E, T, F, \texttt{id}, +, *, (,)\}, \{\texttt{id}, +, *, (,)\}, R, E\}$, where:

$$\begin{aligned} R = \{ & E \rightarrow E + T \\ & E \rightarrow T \\ & T \rightarrow T * F \\ & T \rightarrow F \\ & F \rightarrow (E) \\ & F \rightarrow \texttt{id}\}. \end{aligned}$$

Just as we did for Bal, we have forced branching to go in a single direction (to the left) when identical operators are involved. And, by adding the levels T (for term) and F (for factor) we have defined a precedence hierarchy: Times has

higher precedence than plus does. Using G', there is now a single parse for the string `id + id * id`:



Ambiguous Attachment

The third source of ambiguity that we will consider arises when constructs with optional fragments are nested. The problem in such cases is then, "Given an instance of the optional fragment, at what level of the parse tree should it be attached?"

Probably the most often described instance of this kind of ambiguity is known as the ***dangling else problem***. Suppose that we define a programming language with an `if` statement that can have either of the following forms:

<stmt> ::= `if` <cond> `then` <stmt>

<stmt> ::= `if` <cond> `then` <stmt> `else` <stmt>

In other words, the `else` clause is optional. Then the following statement, with just a single `else` clause, has two parses:

`if` $cond_1$ `then if` $cond_2$ `then` st_1 `else` st_2

In the first parse, the single `else` clause goes with the first `if`. (So it attaches high in the parse tree.) In the second parse, the single `else` clause goes with the second `if`. (In this case, it attaches lower in the parse tree.)

EXAMPLE 11.20 The Dangling Else Problem in Java

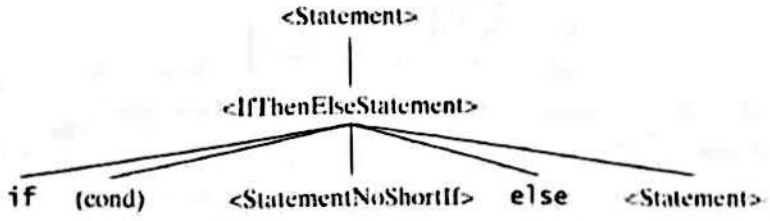
Most programming languages that have the dangling else problem (including C, C++, and Java) specify that each `else` goes with the innermost `if` to which it can be attached. The Java grammar forces this to happen by changing the rules to something like these (presented here in a simplified form that omits many of the statement types that are allowed):

<Statement> ::= <IfThenStatement> | <IfThenElseStatement> |
<IfThenElseStatementNoShortIf> | ...
<StatementNoShortIf> ::= <block> | <IfThenElseStatementNoShortIf> | ...
<IfThenStatement> ::= `if` (<Expression>) <Statement>
<IfThenElseStatement> ::= `if` (<Expression>) <StatementNoShortIf> `else`
<Statement>

EXAMPLE 11.20 (*Continued*)

<IfThenElseStatementNoShortIf> ::= `if` (<Expression>)
<StatementNoShortIf> `else` <StatementNoShortIf>

In this grammar, there is a special class of statements called <Statement NoShortIf>. These are statements that are guaranteed not to end with a short (i.e., `else`-less `if` statement). The grammar uses this class to guarantee that, if a top-level `if` statement has an `else` clause, then any embedded `if` must also have one. To see how this works, consider the following parse tree:



The top-level `if` statement claims the `else` clause for itself by guaranteeing that there will not be an embedded `if` that is missing an `else`. If there were, then that embedded `if` would grab the one `else` clause there is.

> For a discussion of other ways in which programming languages can solve this problem, see G.3.

Attachment ambiguity is also a problem for parsers for natural languages such as English, as we'll see in Example 11.22

Proving that a Grammar is Unambiguous

While it is undecidable, *in general*, whether a grammar is ambiguous or unambiguous, it may be possible to prove that a *particular* grammar is either ambiguous or unambiguous. A grammar G can be shown to be ambiguous by exhibiting a single string for which G produces two parse trees. To see how it might be possible to prove that G is unambiguous, recall that G is unambiguous iff every string derivable in G has a single leftmost derivation. So, if we can show that, during any leftmost derivation of any string $w \in L(G)$, exactly one rule can be applied, then G is unambiguous.

EXAMPLE 11.21 The Final Balanced Parens Grammar is Unambiguous

We return to the final grammar G that we produced for Bal. $G = (\{S,), (, \{, \}, \}, R, S)$, where:

$$
\begin{aligned}
R = \{ & S^* \rightarrow \varepsilon \quad (1) \\
& S^* \rightarrow S \quad (2) \\
& S \rightarrow SS_1 \quad (3) \\
& S \rightarrow S_1 \quad (4) \\
& S_1 \rightarrow (S) \quad (5) \\
& S_1 \rightarrow ()\}. \quad (6)
\end{aligned}
$$

We prove that G is unambiguous. Given the leftmost derivation of any string w in $L(G)$, there is, at each step of the derivation, a unique symbol, which we'll call X, that is the leftmost nonterminal in the working string. Whatever X is, it must be expanded by the next rule application, so the only rules that may be applied next are those with X on the left-hand side. There are three nonterminals in G. We show, for each of them, that the rules that expand them never compete in the leftmost derivation of a particular string w. We do the two easy cases first:

- S^*: The only place that S^* may occur in a derivation is at the beginning. If $w = \varepsilon$, then rule (1) is the only one that can be applied. If $w \neq \varepsilon$, then rule (2) is the only one that can be applied.
- S_1: If the next two characters to be derived are (), S_1 must expand by rule (6). Otherwise, it must expand by rule (5).

In order discuss S, we first define, for any matched set of parentheses m, the ***siblings*** of m to be the smallest set that includes any matched set p adjacent, on the right, to m and all of p's siblings. So, for example, consider the string:

$$\frac{\left(\frac{()}{1}\,\frac{()}{2}\right)\,\frac{()}{3}\,\frac{()}{4}}{5}$$

The set () labeled 1 has a single sibling, 2. The set (()()) labeled 5 has two siblings, 3 and 4. Now we can consider S. We observe that:

- S must generate a string in Bal and so it must generate a matched set, possibly with siblings.
- So the first terminal character in any string that S generates is (. Call the string that starts with that (and ends with the) that matches it, s.
- The only thing that S_1 can generate is a single matched set of parentheses that has no siblings.
- Let n be the number of siblings of s. In order to generate those siblings, S must expand by rule (3) exactly n times (producing n copies of S_1) before it expands by rule (4) to produce a single S_1, which will produce s. So, at every step in a derivation, let p be the number of occurrences of S_1 to the right of S. If $p < n$, S must expand by rule (3). If $p = n$, S must expand by rule (4).

Going Too Far

We must be careful, in getting rid of ambiguity, that we don't do so at the expense of being able to generate the parse trees that we want. In both the arithmetic expression example and the dangling else case, we were willing to force one interpretation. Sometimes, however, that is not an acceptable solution.

EXAMPLE 11.22 Throwing Away The Parses That We Want

Let's return to the small English grammar that we showed in Example 11.6. That grammar is ambiguous. It has an ambiguous attachment problem, similar to the dangling else problem. Consider the following two sentences:

```
Chris likes the girl with a cat.

Chris shot the bear with a rifle.
```

Each of these sentences has two parse trees because, in each case, the prepositional phrase `with a` *N*, can be attached either to the immediately preceding *NP* (`the girl` or `the bear`) or to the *VP*. The correct interpretation for the first sentence is that there is a girl with a cat and Chris likes her. In other words, the prepositional phrase attaches to the *NP*. Almost certainly, the correct interpretation for the second sentence is that there is a bear (with no rifle) and Chris used a rifle to shoot it. In other words, the prepositional phrase attaches to the *VP*. See L.3.4 for additional discussion of this example.

For now, the key point is that we could solve the ambiguity problem by eliminating one of the choices for *PP* attachment. But then, for one of our two sentences, we'd get a parse tree that corresponds to nonsense. In other words, we might still have a grammar with the required weak generative capacity, but we would no longer have one with the required strong generative capacity. The solution to this problem is to add some additional mechanism to the context-free framework. That mechanism must be able to choose the parse that corresponds to the most likely meaning.

> English parsers must have ways to handle various kinds of attachment ambiguities, including those caused by prepositional phrases and relative clauses. (L.3.4)

11.8 Normal Forms ●

So far, we've imposed no restrictions on the form of the right-hand sides of our grammar rules, although we have seen that some kinds of rules, like those whose right-hand side is ε, can make grammars harder to use. In this section, we consider what happens if we carry the idea of getting rid of ε-productions a few steps farther.

> Normal forms for queries and data can simplify database processing. (H.5) Normal forms for logical formulas can simplify automated reasoning in artificial intelligence systems (M.2) and in program verification systems. (H.1.1)

Let C be any set of data objects. For example, C might be the set of context-free grammars. Or it could be the set of syntactically valid logical expressions or a set of database queries. We'll say that a set F is a ***normal form*** for C iff it possesses the following two properties:

- For every element c of C, except possibly a finite set of special cases, there exists some element f of F such that f is equivalent to c with respect to some set of tasks.
- F is simpler than the original form in which the elements of C are written. By "simpler" we mean that at least some tasks are easier to perform on elements of F than they would be on elements of C.

We define normal forms in order to make other tasks easier. For example, it might be easier to build a parser if we could make some assumptions about the form of the grammar rules that the parser will use. Recall that, in Section 5.8, we introduced the notion of a canonical form for a set of objects. A normal form is a weaker notion, since it does not require that there be a unique representation for each object in C, nor does it require that "equivalent" objects map to the same representation. So it is sometimes possible to define useful normal forms when no useful canonical form exists. We'll now do that for context-free grammars.

11.8.1 Normal Forms for Grammars

We'll define the following two useful normal forms for context-free grammars:

- ***Chomsky Normal Form***: In a Chomsky normal form grammar $G = (V, \Sigma, R, S)$, all rules have one of the following two forms:
  - $X \rightarrow a$, where $a \in \Sigma$, or
  - $X \rightarrow BC$, where B and C are elements of $V - \Sigma$.

  Every parse tree that is generated by a grammar in Chomsky normal form has a branching factor of exactly 2, except at the branches that lead to the terminal nodes, where the branching factor is 1. This property makes Chomsky normal form grammars useful in several ways, including:
  - Parsers can exploit efficient data structures for storing and manipulating binary trees.
  - Every derivation of a string w contains $|w| - 1$ applications of some rule of the form $X \rightarrow BC$, and $|w|$ applications of some rule of the form $X \rightarrow a$. So it is straightforward to define a decision procedure to determine whether w can be generated by a Chomsky normal form grammar G.

In addition, because the form of all the rules is so restricted, it is easier than it would otherwise be to define other algorithms that manipulate grammars.

- ***Greibach Normal Form***: In a Greibach normal form grammar $G = (V, \Sigma, R, S)$, all rules have the following form:
  - $X \rightarrow a\beta$, where $a \in \Sigma$ and $\beta \in (V - \Sigma)^*$.

  In every derivation that is produced by a grammar in Greibach normal form, precisely one terminal is generated for each rule application. This property is useful in several ways, including:
  - Every derivation of a string w contains $|w|$ rule applications. So again it is straightforward to define a decision procedure to determine whether w can be generated by a Greibach normal form grammar G.
  - As we'll see in Theorem 14.2, Greibach normal form grammars can easily be converted to pushdown automata with no ε-transitions. This is useful because such PDAs are guaranteed to halt.

THEOREM 11.1 Chomsky Normal Form

Theorem: Given a context-free grammar G, there exists a Chomsky normal form grammar G_C such that $L(G_C) = L(G) - \{\varepsilon\}$.

Proof: The proof is by construction, using the algorithm *convertToChomsky* presented below.

THEOREM 11.2 Greibach Normal Form

Theorem: Given a context-free grammar G, there exists a Greibach normal form grammar G_G such that $L(G_G) = L(G) - \{\varepsilon\}$.

Proof: The proof is also by construction. We present it in D.1.

11.8.2 Converting to a Normal Form

Normal forms are useful if there exists a procedure for converting an arbitrary object into a corresponding object that meets the requirements of the normal form. Algorithms to convert grammars into normal forms generally begin with a grammar G and then operate in a series of steps as follows:

1. Apply some transformation to G to get rid of undesirable property 1. Show that the language generated by G is unchanged.
2. Apply another transformation to G to get rid of undesirable property 2. Show that the language generated by G is unchanged *and* that undesirable property 1 has not been reintroduced.
3. Continue until the grammar is in the desired form.

Because it is possible for one transformation to undo the work of an earlier one, the order in which the transformation steps are performed is often critical to the correctness of the transformation algorithm.

One transformation that we will exploit in converting grammars both to Chomsky normal form and to Greibach normal form is based on the following observation. Consider a grammar that contains the three rules:

$X \rightarrow aYc$
$Y \rightarrow b$
$Y \rightarrow ZZ$

We can construct an equivalent grammar by replacing the X rule with the rules:

$X \rightarrow abc$
$X \rightarrow aZZc$

Instead of letting X generate an instance of Y, X immediately generates whatever Y could have generated. The following theorem generalizes this claim.

THEOREM 11.3 Rule Substitution

Theorem: Let $G = (V, \Sigma, R, S)$ be a context-free grammar that contains a rule r of the form $X \rightarrow \alpha Y \beta$, where α and β are elements of V^* and $Y \in (V - \Sigma)$. Let $Y \rightarrow \gamma_1 | \gamma_2 | \ldots | \gamma_n$ be all of G's rules whose left-hand side is Y. And let G' be the result of removing from R the rule r and replacing it by the rules $X \rightarrow \alpha\gamma_1\beta, X \rightarrow \alpha\gamma_2\beta, \ldots, X \rightarrow \alpha\gamma_n\beta$. Then $L(G') = L(G)$.

Proof: We first show that every string in $L(G)$ is also in $L(G')$: Suppose that w is in $L(G)$. If G can derive w without using rule r, then G' can do so in exactly the same way. If G can derive w using rule r, then one of its derivations has the following form, for some value of k between 1 and n:

$$S \Rightarrow \ldots \Rightarrow \delta X \phi \Rightarrow \delta\alpha Y \beta\phi \Rightarrow \delta\alpha\gamma_k\beta\phi \Rightarrow \ldots \Rightarrow w.$$

Then G' can derive w with the derivation:

$$S \Rightarrow \ldots \Rightarrow \delta X \phi \Rightarrow \quad \delta\alpha\gamma_k\beta\phi \Rightarrow \ldots \Rightarrow w.$$

Next we show that only strings in $L(G)$ can be in $L(G')$. This must be so because the action of every new rule $X \rightarrow \alpha\gamma_k\beta$ could have been performed in G by applying the rule $X \rightarrow \alpha Y \beta$ and then the rule $Y \rightarrow \gamma_k$.

11.8.3 Converting to Chomsky Normal Form

There exists a straightforward four-step algorithm that converts a grammar $G = (V, \Sigma, R, S)$ into a new grammar G_C such that G_C is in Chomsky normal form and $L(G_C) = L(G) - \{\varepsilon\}$. Define:

convertToChomsky(G: CFG) =

1. Let G_C be the result of removing from G all ε-rules, using the algorithm *removeEps*, defined in Section 11.7.4.
2. Let G_C be the result of removing from G_C all unit productions (rules of the form $A \rightarrow B$), using the algorithm *removeUnits* defined below. It is important that *removeUnits* run after ...

productions. Once this step has been completed, all rules whose right-hand sides have length 1 are in Chomsky normal form (i.e., they are composed of a single terminal symbol).

3. Let G_C be the result of removing from G_C all rules whose right-hand sides have length greater than 1 and include a terminal (e.g., $A \rightarrow aB$ or $A \rightarrow BaC$). This step is simple and can be performed by the algorithm *removeMixed* given below. Once this step has been completed, all rules whose right-hand sides have length 1 or 2 are in Chomsky normal form.
4. Let G_C be the result of removing from G_C all rules whose right-hand sides have length greater than 2 (e.g., $A \rightarrow BCDE$). This step too is simple. It can be performed by the algorithm *removeLong* given below.
5. Return G_C.

A ***unit production*** is a rule whose right-hand side consists of a single nonterminal symbol. The job of *removeUnits* is to remove all unit productions and to replace them by a set of other rules that accomplish the job previously done by the unit productions. So, for example, suppose that we start with a grammar G that contains the following rules:

$S \rightarrow XY$
$X \rightarrow A$
$A \rightarrow B \mid a$
$B \rightarrow b$

Once we get rid of unit productions, it will no longer be possible for X to become A (and then B) and thus to go on to generate a or b. So X will need the ability to go directly to a and b, without any intermediate steps. We can define *removeUnits* as follows:

removeUnits(G: CFG) =

1. Let $G' = G$.
2. Until no unit productions remain in G' do:
   2.1. Choose some unit production $X \rightarrow Y$.
   2.2. Remove it from G'.
   2.3. Consider only rules that still remain in G'. For every rule $Y \rightarrow \beta$, where $\beta \in V^*$, do:
   Add to G' the rule $X \rightarrow \beta$ unless that is a rule that has already been removed once.
3. Return G'.

Notice that we have not bothered to check to make sure that we don't insert a rule that is already present. Since R, the set of rules, is a set, inserting an element that is already in the set has no effect.

At each step of its operation, *removeUnits* is performing the kind of rule substitution described in Theorem 11.3. (It happens that both α and β are empty.) So that theorem tells us that, at each step, the language generated by G' is unchanged from the previous step. If *removeUnits* halts, it is clear that all unit productions have been removed. It is less obvious that *removeUnits* can be guaranteed to halt. At each step, one unit production is removed, but several new rules may be added, including new unit productions. To see that *removeUnit* must halt, we observe that there is a bound $= |V - \Sigma|^2$ on the

number of unit productions that can be formed from a fixed set $V - \Sigma$ of nonterminals. At each step, *removeUnits* removes one element from that set and that element can never be reinserted. So *removeUnits* must halt in at most $|V - \Sigma|^2$ steps.

EXAMPLE 11.23 Removing Unit Productions

Let $G = (V, \Sigma, R, S)$, where:

$$R = \{S \rightarrow XY$$
$$X \rightarrow A$$
$$A \rightarrow B \mid \mathrm{a}$$
$$B \rightarrow \mathrm{b}$$
$$Y \rightarrow T$$
$$T \rightarrow Y \mid \mathrm{c}\}.$$

The order in which *removeUnits* chooses unit productions to remove doesn't matter. We'll consider one order it could choose:

Remove $X \rightarrow A$. Since $A \rightarrow B \mid \mathrm{a}$, add $X \rightarrow B \mid \mathrm{a}$.
Remove $X \rightarrow B$. Add $X \rightarrow \mathrm{b}$.
Remove $Y \rightarrow T$. Add $Y \rightarrow Y \mid \mathrm{c}$. Notice that we've added $Y \rightarrow Y$, which is useless, but it will be removed later.
Remove $Y \rightarrow Y$. Consider adding $Y \rightarrow T$, but don't since it has previously been removed.
Remove $A \rightarrow B$. Add $A \rightarrow \mathrm{b}$.
Remove $T \rightarrow Y$. Add $T \rightarrow \mathrm{c}$, but with no effect since it was already present.

At this point, the rules of G are:

$$S \rightarrow XY$$
$$A \rightarrow \mathrm{a} \mid \mathrm{b}$$
$$B \rightarrow \mathrm{b}$$
$$T \rightarrow \mathrm{c}$$
$$X \rightarrow \mathrm{a} \mid \mathrm{b}$$
$$Y \rightarrow \mathrm{c}$$

No unit productions remain, so *removeUnits* halts.

We must now define the two straightforward algorithms that are required by steps 3 and 4 of the conversion algorithm that we sketched above. We begin by defining:

removeMixed (G: CFG) =

1. Let $G' = G$.
2. Create a new nonterminal T_a for each terminal a in Σ.
3. Modify each rule in G' whose right-hand side has length greater than 1 and that contains a terminal symbol by substituting T_a for each occurrence of the terminal a.
4. Add to G', for each T_a, the rule $T_a \rightarrow a$.
5. Return G'.

EXAMPLE 11.24 Removing Mixed Productions

The result of applying *removeMixed* to the grammar:

$A \rightarrow \mathrm{a}$
$A \rightarrow \mathrm{a}B$
$A \rightarrow B\mathrm{a}C$
$A \rightarrow B\mathrm{b}C$

is the grammar:

$A \rightarrow \mathrm{a}$
$A \rightarrow T_a B$
$A \rightarrow B T_a C$
$A \rightarrow B T_b C$
$T_a \rightarrow \mathrm{a}$
$T_b \rightarrow \mathrm{b}$

Finally we define *removeLong*. The idea for *removeLong* is simple. If there is a rule with n symbols on its right-hand side, replace it with a set of rules. The first rule generates the first symbol followed by a new symbol that will correspond to "the rest". The next rule rewrites that symbol as the second of the original symbols, followed by yet another new one, again corresponding to "the rest", and so forth, until there are only two symbols left to generate. So we define:

removeLong (G: CFG) =

1. Let $G' = G$.
2. For each G' rule r^k of the form $A \rightarrow N_1 N_2 N_3 N_4 \ldots N_n$, $n > 2$, create new nonterminals $M^k_2, M^k_3, \ldots M^k_{n-1}$.
3. In G', replace r^k with the rule $A \rightarrow N_1 M^k_2$.
4. To G', add the rules $M^k_2 \rightarrow N_2 M^k_3$, $M^k_3 \rightarrow N_3 M^k_4, \ldots M^k_{n-1} \rightarrow N_{n-1} N_n$.
5. Return G'.

When we illustrate this algorithm, we typically omit the superscripts on the M's, and, instead, guarantee that we use distinct nonterminals by using distinct subscripts.

EXAMPLE 11.25 Removing Rules with Long Right-hand Sides

The result of applying *removeLong* to the single rule grammar:

$$A \rightarrow BCDEF$$

is the grammar with rules:

$A \rightarrow BM_2$
$M_2 \rightarrow CM_3$
$M_3 \rightarrow DM_4$
$M_4 \rightarrow EF$

We can now illustrate the four steps of *converttoChomsky*.

EXAMPLE 11.26 Converting a Grammar to Chomsky Normal Form

Let $G = (\{S, A, B, C, \mathrm{a}, \mathrm{c}\}, \{A, B, C\}, R, S)$, where:

$$\begin{aligned} R = \{ & S \rightarrow \mathrm{a}AC\mathrm{a} \\ & A \rightarrow B \mid \mathrm{a} \\ & B \rightarrow C \mid \mathrm{c} \\ & C \rightarrow \mathrm{c}C \mid \varepsilon\}. \end{aligned}$$

We convert G to Chomsky normal form. Step 1 applies *removeEps* to eliminate ε-productions. We compute N, the set of nullable variables. Initially $N = \{C\}$. Because of the rule $B \rightarrow C$, we add B. Then, because of the rule $A \rightarrow B$, we add A. So $N = \{A, B, C\}$. Since both A and C are nullable, we derive three new rules from the first original rule, giving us:

$$S \rightarrow \mathrm{a}AC\mathrm{a} \mid \mathrm{a}A\mathrm{a} \mid \mathrm{a}C\mathrm{a} \mid \mathrm{aa}$$

We add $A \rightarrow \varepsilon$ and $B \rightarrow \varepsilon$, but both of them will disappear at the end of this step. We also add $C \rightarrow \mathrm{c}$. So *removeEps* returns the rule set:

$$\begin{aligned} & S \rightarrow \mathrm{a}AC\mathrm{a} \mid \mathrm{a}A\mathrm{a} \mid \mathrm{a}C\mathrm{a} \mid \mathrm{aa} \\ & A \rightarrow B \mid \mathrm{a} \\ & B \rightarrow C \mid \mathrm{c} \\ & C \rightarrow \mathrm{c}C \mid \mathrm{c} \end{aligned}$$

Next we apply *removeUnits*:

Remove $A \rightarrow B$. Add $A \rightarrow C \mid \mathrm{c}$.
Remove $B \rightarrow C$. Add $B \rightarrow \mathrm{c}C$ (and $B \rightarrow \mathrm{c}$, but it was already there).
Remove $A \rightarrow C$. Add $A \rightarrow \mathrm{c}C$ (and $A \rightarrow \mathrm{c}$, but it was already there).

So *removeUnits* returns the rule set:

$$\begin{aligned} & S \rightarrow \mathrm{a}AC\mathrm{a} \mid \mathrm{a}A\mathrm{a} \mid \mathrm{a}C\mathrm{a} \mid \mathrm{aa} \\ & A \rightarrow \mathrm{a} \mid \mathrm{c} \mid \mathrm{c}C \\ & B \rightarrow \mathrm{c} \mid \mathrm{c}C \\ & C \rightarrow \mathrm{c}C \mid \mathrm{c} \end{aligned}$$

Next we apply *removeMixed*, which returns the rule set:

$$\begin{aligned} & S \rightarrow T_a A C T_a \mid T_a A T_a \mid T_a C T_a \mid T_a T_a \\ & A \rightarrow \mathrm{a} \mid \mathrm{c} \mid T_c C \\ & B \rightarrow \mathrm{c} \mid T_c C \\ & C \rightarrow T_c C \mid \mathrm{c} \end{aligned}$$

EXAMPLE 11.26 (*Continued*)

$T_a \rightarrow a$
$T_c \rightarrow c$

Finally, we apply *removeLong*, which returns the rule set:

$S \rightarrow T_a S_1 \quad S \rightarrow T_a S_3 \quad S \rightarrow T_a S_4 \quad S \rightarrow T_a T_a$
$S_1 \rightarrow A S_2 \quad S_3 \rightarrow A T_a \quad S_4 \rightarrow C T_a$
$S_2 \rightarrow C T_a$
$A \rightarrow a \mid c \mid T_c C$
$B \rightarrow c \mid T_c C$
$C \rightarrow T_c C \mid c$
$T_a \rightarrow a$
$T_c \rightarrow c$

From Example 11.26 we see that the Chomsky normal form version of a grammar may be longer than the original grammar was. How much longer? And how much time may be required to execute the conversion algorithm? We can answer both of these questions by answering them for each of the steps that the conversion algorithm executes. Let n be the length of an original grammar G. Then we have:

1. Use *removeEps* to remove ε-rules: Suppose that G contains a rule of the form $X \rightarrow A_1 A_2 A_3 \ldots A_k$. If all of the variables A_1 through A_k are nullable, this single rule will be rewritten as $2^k - 1$ rules (since each of the k nonterminals can either be present or not, except that they cannot all be absent). Since k can grow as n, we have that the length of the grammar that *removeEps* produces (and thus the amount of time that *removeEps* requires) is $\mathcal{O}(2^n)$ In this worst case, the conversion algorithm becomes impractical for all but toy grammars. We can prevent this worst case from occurring though. Suppose that all right-hand sides can be guaranteed to be short. For example, suppose they all have length at most 2. Then no rule will be rewritten as more than 3 rules. We can make this guarantee if we modify *converttoChomsky* slightly. We will run *removeLong* as step 1 rather than as step 4. Note that none of the other steps can create a rule whose right-hand side is longer than the right-hand side of some rule that already exists. So it is not necessary to rerun *removeLong* later. With this change, *removeEps* runs in linear time.
2. Use *removeUnits* to remove unit productions: We've already shown that this step must halt in at most $|V - \Sigma|^2$ steps. Each of those steps takes constant time and may create one new rule. So the length of the grammar that *removeUnits* produces, as well as the time required for it to run, is $\mathcal{O}(n^2)$.
3. Use *removeMixed* to remove rules with right-hand sides of length greater than 1 and that contain a terminal symbol: This step runs in linear time and constructs a grammar whose size grows linearly.

4. Use *removeLong* to remove rules with long right-hand sides: This step runs in linear time and constructs a grammar whose size grows linearly.

So, if we change *converttoChomsky* so that it does step 4 first, its time complexity is $\mathcal{O}(n^2)$ and the size of the grammar that it produces is also $\mathcal{O}(n^2)$.

11.8.4 The Price of Normal Forms

While normal forms are useful for many things, as we will see over the next few chapters, it is important to keep in mind that they exact a price and it's one that we may or may not be willing to pay, depending on the application. If G is an arbitrary context-free grammar and G' is an equivalent grammar in Chomsky (or Greibach) normal form, then G and G' generate the same set of strings, but only in rare cases (for example if G happened already to be in normal form) do they assign to those strings the same parse trees. Thus, while converting a grammar to a normal form has no effect on its weak generative capacity, it may have a significant effect on its strong generative capacity.

11.9 Island Grammars ●

Suppose that we want to parse strings that possess one or more of the following properties:

- Some (perhaps many) of them are ill-formed. In other words, while there may be a grammar that describes what strings are "supposed to look like", there is no guarantee that the actual strings we'll see conform to those rules. Consider, for example, any grammar you can imagine for English. Now imagine picking up the phone and hearing something like, "Um, I uh need a copy of uh my bill for er Ap, no May, I think, or June, maybe all of them uh, I guess that would work." Or consider a grammar for HTML. It will require that tags be properly nested. But strings like `<b><i>bold italic</b></i>` show up not infrequently in HTML documents. Most browsers will do the right thing with them, so they never get debugged.
- We simply don't know enough about them to build an exact model, although we do know something about some patterns that we think the strings will contain.
- They may contain substrings in more than one language. For example, bi(multi)lingual people often mix their speech. We even give names to some of the resulting hybrids: Spanglish, Japlish, Hinglish, etc. Or consider a typical Web page. It may contain fragments of HTML, Java script, or other languages, interleaved with each other. Even when parsing strings that are all in the same "language", dialectical issues may arise. For example, in response to the question, "Are you going to fix dinner tonight?" an American speaker of English might say, "I could," while a British speaker of English might say, "I could do." Similarly, in analyzing legacy software, there are countless dialects of languages like Fortran and Cobol.
- They may contain some substrings we care about, interleaved with other substrings we don't care about and don't want to waste time parsing. For example, when parsing an XML document to determine its top level structure, we may have no interest in the text or even in many of the tags.

Then the sentence s that is most likely to have been generated, given the observation o, is the one with the highest conditional probability given o. Recall that argmax of w returns the value of the argument w that maximizes the value of the function it is given. So the highest probability sentence s is:

$$s = \underset{w \in X}{\text{argmax}}\, Pr\,(w|o)$$

$$= \underset{w \in X}{\text{argmax}} \frac{\Pr\,(o|w)\Pr\,(w)}{\Pr\,(o)}.$$

Stochastic context-free grammars can be used model the three-dimensional structure of RNA. (K.4)

In Chapter 15, we will discuss techniques for parsing context-free languages that are defined by standard (i.e., without probabilistic information) context-free grammars. Those techniques can be extended to create techniques for parsing using stochastic grammars. So they can be used to answer both of the questions that we just presented.

Exercises

1. Let $\Sigma = \{a, b\}$. For the languages that are defined by each of the following grammars, do each of the following:
   i. List five strings that are in L.
   ii. List five strings that are not in L (or as many as there are, whichever is greater).
   iii. Describe L concisely. You can use regular expressions, expressions using variables (e.g., $a^n b^n$, or set theoretic expressions (e.g., $\{x: \ldots\}$).
   iv. Indicate whether or not L is regular. Prove your answer.

   a. $S \rightarrow aS \mid Sb \mid \varepsilon$
   b. $S \rightarrow aSa \mid bSb \mid a \mid b$
   c. $S \rightarrow aS \mid bS \mid \varepsilon$
   d. $S \rightarrow aS \mid aSbS \mid \varepsilon$
2. Let G be the grammar of Example 11.12. Show a third parse tree that G can produce for the string (())().
3. Consider the following grammar G:

   $$S \rightarrow 0S1|SS|10$$

   Show a parse tree produced by G for each of the following strings:
   a. 010110.
   b. 00101101.
4. Consider the following context free grammar G:

   $$S \rightarrow aSa$$

$$S \rightarrow T$$
$$S \rightarrow \varepsilon$$
$$T \rightarrow \mathsf{b}T$$
$$T \rightarrow \mathsf{c}T$$
$$T \rightarrow \varepsilon$$

One of these rules is redundant and could be removed without altering $L(G)$. Which one?

5. Using the simple English grammar that we showed in Example 11.6, show two parse trees for each of the following sentences. In each case, indicate which parse tree almost certainly corresponds to the intended meaning of the sentence:
   a. `The bear shot Fluffy with the rifle.`
   b. `Fluffy likes the girl with the chocolate.`
6. Show a context-free grammar for each of the following languages L:
   a. BalDelim = $\{w$: where w is a string of delimiters: (,), [,], {, }, that are properly balanced$\}$.
   b. $\{\mathsf{a}^i\mathsf{b}^j : 2i = 3j + 1\}$.
   c. $\{\mathsf{a}^i\mathsf{b}^j : 2i \neq 3j + 1\}$.
   d. $\{w \in \{\mathsf{a}, \mathsf{b}\}^* : \#_{\mathsf{a}}(w) = 2 \cdot \#_{\mathsf{b}}(w)\}.\}$.
   e. $L = \{w \in \{\mathsf{a}, \mathsf{b}\}^* : w = w^{\mathrm{R}}\}$.
   f. $\{\mathsf{a}^i\mathsf{b}^j\mathsf{c}^k : i, j, k \geq 0$ and $(i \neq j$ or $j \neq k)\}$.
   g. $\{\mathsf{a}^i\mathsf{b}^j\mathsf{c}^k : i, j, k \geq 0$ and $(k \leq i$ or $k \leq j)\}$.
   h. $\{w \in \{\mathsf{a}, \mathsf{b}\}^*$: every prefix of w has at least as many a's as b's$\}$.
   i. $\{\mathsf{a}^n\mathsf{b}^m : m \geq n, m\text{-}n$ is even$\}$.
   j. $\{\mathsf{a}^m\mathsf{b}^n\mathsf{c}^p\mathsf{d}^q : m, n, p, q \geq 0$ and $m + n = p + q\}$.
   k. $\{x\mathsf{c}^n : x \in \{\mathsf{a}, \mathsf{b}\}^*$ and $(\#_{\mathsf{a}}(x) = n$ or $\#_{\mathsf{b}}(x) = n)\}$.
   l. $\{b_i\#b_{i+1}{}^{\mathrm{R}} : b_i$ is the binary representation of some integer i, $i \geq 0$, without leading zeros$\}$. (For example 101#011 $\in L$.)
   m. $\{x^{\mathrm{R}}\#y : x, y \in \{0, 1\}^*$ and x is a substring of $y\}$.
7. Let G be the ambiguous expression grammar of Example 11.14. Show at least three different parse trees that can be generated from G for the string `id+id*id*id`.
8. Consider the unambiguous expression grammar G' of Example 11.19.
   a. Trace a derivation of the string `id+id*id*id` in G'.
   b. Add exponentiation (**) and unary minus (−) to G', assigning the highest precedence to unary minus, followed by exponentiation, multiplication, and addition, in that order.
9. Let $L = \{w \in \{\mathsf{a}, \mathsf{b}, \cup, \varepsilon, (,), *, ^+\}^* : w$ is a syntactically legal regular expression$\}$.
   a. Write an unambiguous context-free grammar that generates L. Your grammar should have a structure similar to the arithmetic expression grammar G' that we presented in Example 11.19. It should create parse trees that:

- Associate left given operators of equal precedence, and
- Correspond to assigning the following precedence levels to the operators (from highest to lowest):
  - * and $^+$
  - concatenation
  - $\cup$

b. Show the parse tree that your grammar will produce for the string (a $\cup$ b) ba*.

10. Let $L = \{w \in \{A - Z, \neg, \wedge, \vee, \rightarrow, (,)\}^* : w$ is a syntactically legal Boolean expression$\}$.

a. Write an unambiguous context-free grammar that generates L and that creates parse trees that:

- Associate left given operators of equal precedence, and
- Correspond to assigning the following precedence levels to the operators (from highest to lowest): $\neg$, $\wedge$, $\vee$, and $\rightarrow$.

b. Show the parse tree that your grammar will produce for the string:

$$\neg P \vee R \rightarrow Q \rightarrow S$$

11. In I.3.1, we present a simplified grammar for URIs (Uniform Resource Identifiers), the names that we use to refer to objects on the Web.

a. Using that grammar, show a parse tree for:

`https://www.mystuff.wow/widgets/fradgit#sword`

b. Write a regular expression that is equivalent to the grammar that we present.

12. Prove that each of the following grammars is correct:

a. The grammar, shown in Example 11.3, for the language PalEven.

b. The grammar, shown in Example 11.1, for the language Bal.

13. For each of the following grammars G, show that G is ambiguous. Then find an equivalent grammar that is not ambiguous.

a. $(\{S, A, B, T, \mathrm{a}, \mathrm{c}\}, \{\mathrm{a}, \mathrm{c}\}, R, S)$, where $R = \{S \rightarrow AB, S \rightarrow BA, A \rightarrow \mathrm{a}A, A \rightarrow \mathrm{ac}, B \rightarrow T\mathrm{c}, T \rightarrow \mathrm{a}T, T \rightarrow \mathrm{a}\}$.

b. $(\{S, \mathrm{a}, \mathrm{b}\}, \{\mathrm{a}, \mathrm{b}\}, R, S)$, where $R = \{S \rightarrow \varepsilon, S \rightarrow \mathrm{a}S\mathrm{a}, S \rightarrow \mathrm{b}S\mathrm{b}, S \rightarrow \mathrm{a}S\mathrm{b}, S \rightarrow \mathrm{b}S\mathrm{a}, S \rightarrow SS\}$.

c. $\{\{S, A, B, T, \mathrm{a}, \mathrm{c}\}, \{\mathrm{a}, \mathrm{c}\}, R, S)$, where $R = \{S \rightarrow AB, A \rightarrow AA, A \rightarrow \mathrm{a}, B \rightarrow T\mathrm{c}, T \rightarrow \mathrm{a}T, T \rightarrow \mathrm{a}\}$.

d. $(\{S, \mathrm{a}, \mathrm{b}\}, \{\mathrm{a}, \mathrm{b}\}, R, S)$, where $R = \{S \rightarrow \mathrm{a}S\mathrm{b}, S \rightarrow \mathrm{b}S\mathrm{a}, S \rightarrow SS, S \rightarrow \varepsilon\}$. ($G$ is the grammar that we presented in Example 11.10 for the language $L = \{w \in \{\mathrm{a},\mathrm{b}\}^* : \#_{\mathrm{a}}(w) = \#_{\mathrm{b}}(w)\}$.)

e. $(\{S, \mathrm{a}, \mathrm{b}\}, \{\mathrm{a}, \mathrm{b}\}, R, S)$, where $R = \{S \rightarrow \mathrm{a}S\mathrm{b}, S \rightarrow \mathrm{aa}S\mathrm{b}, S \rightarrow \varepsilon\}$.

14. Let G be any context-free grammar. Show that the number of strings that have a derivation in G of length n or less, for any $n > 0$, is finite.

15. Consider the fragment of a Java grammar that is presented in Example 11.20. How could it be changed to force each `else` clause to be attached to the outermost possible `if` statement?

16. How does the COND form in Lisp, as described in G.5, avoid the dangling else problem?
17. Consider the grammar G' of Example 11.19.
    a. Convert G' to Chomsky normal form.
    b. Consider the string `id*id+id`.
        i. Show the parse tree that G' produces for it.
        ii. Show the parse tree that your Chomsky normal form grammar produces for it.
18. Convert each of the following grammars to Chomsky normal form:
    a. $S \rightarrow \mathtt{a}S\mathtt{a}$
       $S \rightarrow B$
       $B \rightarrow \mathtt{bb}C$
       $B \rightarrow \mathtt{bb}$
       $C \rightarrow \varepsilon$
       $C \rightarrow \mathtt{c}C$
    b. $S \rightarrow ABC$
       $A \rightarrow \mathtt{a}C \mid D$
       $B \rightarrow \mathtt{b}B \mid \varepsilon \mid A$
       $C \rightarrow A\mathtt{c} \mid \varepsilon \mid C\mathtt{c}$
       $D \rightarrow \mathtt{aa}$
    c. $S \rightarrow \mathtt{a}TV\mathtt{a}$
       $T \rightarrow \mathtt{a}T\mathtt{a} \mid \mathtt{b}T\mathtt{b} \mid \varepsilon \mid V$
       $V \rightarrow \mathtt{c}V\mathtt{c} \mid \varepsilon$