

# *Theory of Computation*

Course note based on *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science*, 2nd edition, authored by Martin Davis, Ron Sigal, and Elaine J. Weyuker.

course note prepared by

Tyng–Ruey Chuang

Week 11, Spring 2010

## About This Course Note

- It is prepared for the course *Theory of Computation* taught at the National Taiwan University in Spring 2010.
- It follows very closely the book *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science*, 2nd edition, by Martin Davis, Ron Sigal, and Elaine J. Weyuker. Morgan Kaufmann Publishers. ISBN: 0-12-206382-1.
- It is available from Tyng-Ruey Chuang’s web site:

<http://www.iis.sinica.edu.tw/~trc/>

and released under a Creative Commons “Attribution-ShareAlike 3.0 Taiwan” license:

<http://creativecommons.org/licenses/by-sa/3.0/tw/>

## 1 Context-Free Languages (10)

### 1.1 Closure Properties (10.5)

$L_1 \cup L_2$

**Theorem 5.1.** If  $L_1, L_2$  are context-free languages, then so is  $L_1 \cup L_2$ . *Proof.* Let  $L_1 = L(\Gamma_1), L_2 = L(\Gamma_2)$ , where  $\Gamma_1, \Gamma_2$  are context-free grammars with disjoint sets of variables  $\mathcal{V}_1$  and  $\mathcal{V}_2$ , and start symbols  $S_1, S_2$ , respectively. Let  $\Gamma$  be the context-free grammar with variables  $\mathcal{V}_1 \cup \mathcal{V}_2 \cup \{S\}$  and start symbol  $S$ . The productions of  $\Gamma$  are those of  $\Gamma_1$  and  $\Gamma_2$ , together with the two additional productions  $S \rightarrow S_1$  and  $S \rightarrow S_2$ .

Obviously  $L(\Gamma) = L(\Gamma_1) \cup L(\Gamma_2)$ . □

$L_1 \cap L_2$

**Theorem 5.2.** There are context-free languages  $L_1$  and  $L_2$  such that  $L_1 \cap L_2$  is not context-free. *Proof.* The following two languages  $L_1$  and  $L_2$  are context free.

$$\begin{aligned} L_1 &= \{a^{[n]}b^{[n]}c^{[m]} \mid n, m > 0\} \\ L_2 &= \{a^{[m]}b^{[n]}c^{[n]} \mid n, m > 0\} \end{aligned}$$

However, as shown by Theorem 4.2, their intersection

$$L_1 \cap L_2 = \{a^{[n]}b^{[n]}c^{[n]} \mid n > 0\}$$

is not context-free. □

$A^* - L$

**Corollary 5.3.** There is a context-free language  $L \subseteq A^*$  such that  $A^* - L$  is not context-free. *Proof.* Suppose otherwise, that is, for every context-free language  $L \subseteq A^*$ ,  $A^* - L$  is context-free. Then the De Morgan identity

$$L_1 \cap L_2 = A^* - ((A^* - L_1) \cup (A^* - L_2))$$

together with Theorem 5.1 would contradict Theorem 5.2. □

$R \cap L$

**Theorem 5.4.** If  $R$  is a regular language and  $L$  is a context-free language, then  $R \cap L$  is context-free. *Proof.* Let  $A$  be an alphabet such that  $L, R \in A^*$ . Let  $L = L(\Gamma)$  or  $L(\Gamma) \cup \{0\}$ , where  $\Gamma$  is a positive context-free grammar with variables  $\mathcal{V}$ , terminals  $A$  and start symbol  $S$ . Let  $\mathcal{M}$  be a dfa that accepts  $R$  with states  $Q$ , initial state  $q_1 \in Q$ , accepting states  $F \subseteq Q$ , and transition function  $\delta$ . For each symbol  $\sigma \in A \cup \mathcal{V}$ , and each ordered pair  $p, q \in Q$ , we introduce a new symbol  $\sigma^{pq}$ . We shall construct a positive context-free grammar  $\tilde{\Gamma}$  whose terminals are  $A$ , and whose variables consists of a start symbol  $\tilde{S}$  together with all the new symbols  $\sigma^{pq}$  for  $\sigma \in A \cup \mathcal{V}$  and  $p, q \in Q$ . (Note that for  $a \in A$ ,  $a$  is a terminal, but  $a^{pq}$  is a variable for each  $p, q \in Q$ .)

**$R \cap L$ , Continued**

*Proof of Theorem 5.4 (Continued).* The productions of  $\tilde{\Gamma}$  are:

1.  $\tilde{S} \rightarrow S^{q_1 q}$  for all  $q \in F$ .
2.  $X^{pq} \rightarrow \sigma_1^{pr_1} \sigma_2^{r_1 r_2} \dots \sigma_n^{r_{n-1} q}$  of all productions  $X \rightarrow \sigma_1 \sigma_2 \dots \sigma_n$  of  $\Gamma$  and all  $p, r_1, r_2, \dots, r_{n-1}, q \in Q$ .
3.  $a^{pq} \rightarrow a$  for all  $a \in A$  and all  $p, q \in Q$  such that  $\delta(p, a) = q$ .

We shall now prove that  $L(\tilde{\Gamma}) = R \cap L(\Gamma)$ . First let  $u = a_1 a_2 \dots a_n \in R \cap L(\Gamma)$ . Since  $u \in L(\Gamma)$ , we have  $S \Rightarrow_{\tilde{\Gamma}}^* a_1 a_2 \dots a_n$ . It follows that  $\tilde{S} \Rightarrow_{\tilde{\Gamma}} S^{q_1 q_{n+1}} \Rightarrow_{\tilde{\Gamma}}^* a_1^{q_1 q_2} a_2^{q_2 q_3} \dots a_n^{q_n q_{n+1}}$ , where  $q_1, q_2, \dots, q_n, q_{n+1} \in Q$ ,  $q_1$  is the initial state, and  $q_{n+1} \in F$ . Since  $u \in L(\mathcal{M})$ , we can choose states so that  $\delta(q_i, a_i) = q_{i+1}$ , for all  $i$ . This implies that  $a_i^{q_i q_{i+1}} \rightarrow a_i$ , for all  $i$ . We conclude that  $\tilde{S} \Rightarrow_{\tilde{\Gamma}}^* a_1 a_2 \dots a_n$ , hence  $u \in L(\tilde{\Gamma})$ .

### **$R \cap L$ , Continued**

For the other direction, that if  $\tilde{S} \Rightarrow_{\tilde{\Gamma}} S^{q_1 q} \Rightarrow_{\tilde{\Gamma}}^* a_1 a_2 \dots a_n = u$  where  $q \in F$ , then  $S \Rightarrow_{\Gamma}^* u$ , we need to prove the following lemma. **Lemma.** Let  $\sigma^{pq} \Rightarrow_{\tilde{\Gamma}}^* u \in A^*$ . Then,  $\delta^*(p, u) = q$ . Moreover, if  $\sigma$  is a variable, then  $\sigma \Rightarrow_{\tilde{\Gamma}}^* u$ . Proof of this lemma can be done by an induction on the length of a derivation of  $u$  from  $\sigma^{pq} \in \tilde{\Gamma}$ . That is, for derivation of length  $> 2$ , we can write

$$\sigma^{pq} \Rightarrow_{\tilde{\Gamma}} \sigma_1^{r_0 r_1} \sigma_2^{r_1 r_2} \dots \sigma_n^{r_{n-1} r_n} \Rightarrow_{\tilde{\Gamma}}^* u_1 u_2 \dots u_n = u$$

where  $r_0 = p, r_n = q$ , and  $\sigma_i^{r_{i-1} r_i} \Rightarrow_{\tilde{\Gamma}}^* u_i$ . The induction hypotheses ensure that  $\delta^*(r_{i-1}, u_i) = r_i$  and  $\sigma_i \Rightarrow_{\tilde{\Gamma}}^* u_i$ , for all  $i$ . From this we can show that  $\delta^*(p, u) = q$  and  $\sigma \Rightarrow_{\tilde{\Gamma}}^* u$ , hence complete the proof for the other direction.  $\square$

### **Erased Symbols**

Let  $A, P$  be alphabets such that  $P \subseteq A$ . For each letter  $a \in A$ , let us write

$$a^0 = \begin{cases} 0 & \text{if } a \in P \\ a & \text{if } a \in A - P. \end{cases}$$

If  $x = a_1 a_2 \dots a_n \in A^*$ , we write

$$\text{Er}_P(x) = a_1^0 a_2^0 \dots a_n^0$$

In other words,  $\text{Er}_P(x)$  is the word that results from  $x$  where all the symbols in it that are part of the alphabet  $P$  are “erased.”

### **Erased Symbols, Continued**

If  $L \subseteq A^*$ , we also write

$$\text{Er}_P(L) = \{\text{Er}_P(x) \mid x \in L\}.$$

If  $\Gamma$  is any context-free grammar with terminal symbols  $T$  and if  $P \subseteq T$ , we write  $\text{Er}_P(\Gamma)$  for the context-free grammar with terminals  $T - P$ , the same variables and start symbol as  $\Gamma$ , and production

$$X \rightarrow \text{Er}_P(v)$$

for each production  $X \rightarrow v$  of  $\Gamma$ .

### **A Theorem about Erased Symbols**

**Theorem 5.5.** If  $\Gamma$  is a context-free grammar and  $\tilde{\Gamma} = \text{Er}_P(\Gamma)$ , then  $L(\tilde{\Gamma}) = \text{Er}_P(L(\Gamma))$ . *Proof Outline.* Suppose that  $w \in L(\Gamma)$ , we have

$$S = w_1 \Rightarrow_{\Gamma} w_2 \dots \Rightarrow_{\Gamma} w_m = w.$$

Let  $v_i = \text{Er}_P(w_i), i = 1, 2, \dots, m$ . Clearly,

$$S = v_1 \Rightarrow_{\tilde{\Gamma}} v_2 \dots \Rightarrow_{\tilde{\Gamma}} v_m = \text{Er}_P(w).$$

so that  $\text{Er}_P(w) \in L(\tilde{\Gamma})$ . This proves that  $L(\tilde{\Gamma}) \supseteq \text{Er}_P(L(\Gamma))$ . For the other direction, we need to show that whenever  $X \Rightarrow_{\tilde{\Gamma}}^* v \in (T - P)^*$ , there is a word  $w \in T^*$  such that  $X \Rightarrow_{\Gamma}^* w$  and  $v = \text{Er}_P(w)$ . This can be done by an induction on the length of a derivation of  $v$  from  $X$  in  $\tilde{\Gamma}$ .  $\square$

### A Theorem about Erased Symbols, Continued

From Theorem 5.5, we may say that the “operators”  $L$  and  $\text{Er}_P$  commute

$$L(\text{Er}_P(\Gamma)) = \text{Er}_P(L(\Gamma))$$

for any context-free grammar  $\Gamma$ . We prove the straightforward: **Corollary 5.6.** If  $L \subseteq A^*$  is a context-free language and  $P \subseteq A$ , then  $\text{Er}_P(L)$  is also a context-free language. *Proof.* Let  $L = L(\Gamma)$ , where  $\Gamma$  is context-free grammar. Let  $\tilde{\Gamma} = \text{Er}_P(\Gamma)$ . By Theorem 5.5,  $\text{Er}_P(\Gamma) = L(\tilde{\Gamma})$  so is context-free.  $\square$

## 1.2 Bracket Languages (10.7)

### Bracket Languages

Let  $A$  be a finite set. Let  $B$  be an alphabet we get from  $A$  by adding  $2n$  new symbols  $(i, )_i, i = 1, 2, \dots, n$ , where  $n$  is some given positive integer. We write  $\text{PAR}_n(A)$  for the language consisting of all the strings in  $B^*$  that are correctly “paired,” thinking of each pair  $(i, )_i$  as matching left and right brackets. More precisely,  $\text{PAR}_n(A) = L(\Gamma_0)$ , where  $\Gamma_0$  is the context-free grammar with the single variables  $S$ , terminals  $B$ , and the productions

1.  $S \rightarrow a$  for all  $a \in A$ ,
2.  $S \rightarrow (iS)_i, \quad i = 1, 2, \dots, n$ ,
3.  $S \rightarrow SS, \quad S \rightarrow 0$ .

The languages  $\text{PAR}_n(A)$  are called *bracket languages*.

### Bracket Languages, Examples

Let  $A = \{a, b, c\}$ , and  $n = 2$ . For ease of reading we will use the symbol ( for  $(_1, )$  for  $)_1$ , [ for  $(_2$ , and ] for  $)_2$ . Then we have

$$cb[(ab)c](a[b]c) \in \text{PAR}_2(A)$$

as well as

$$()[] \in \text{PAR}_2(A)$$

## Bracket Languages, Properties

**Theorem 7.1.**  $\text{PAR}_n(A)$  is a context-free language such that

1.  $A^* \subseteq \text{PAR}_n(A)$ ;
2. if  $x, y \in \text{PAR}_n(A)$ , so is  $xy$ ;
3. if  $x \in \text{PAR}_n(A)$ , so is  $(ix)_i$ , for  $i = 1, 2, \dots, n$ ;
4. if  $x \in \text{PAR}_n(A)$  and  $x \notin A^*$ , then we can write  $x = u(i v)_i w$ , for some  $i = 1, 2, \dots, n$ , where  $u \in A^*$  and  $v, w \in \text{PAR}_n(A)$ .

*Proof Outline.* The proof for the first three properties are straightforward. For the last, we use an induction on the length of  $x$ . Note we have  $|x| > 1$  otherwise  $x \in A \subseteq A^*$ , a contradiction.

Since  $|x| > 1$ , we need only to consider two cases:

- $S \Rightarrow (iS)_i \Rightarrow^* (i v)_i = x$ , where  $S \Rightarrow^* v$ ;
- $S \Rightarrow SS \Rightarrow^* rs = x$ , where  $S \Rightarrow^* r, S \Rightarrow^* s$ , and  $r \neq \emptyset, s \neq \emptyset$ .

Both lead to  $x = u(i v)_i w$ ,  $u \in A^*$  and  $v, w \in \text{PAR}_n(A)$ . □

## Dyck Languages

The language  $\text{PAR}_n(\emptyset)$  is called the *Dyck language* of order  $n$  and is usually written  $D_n$ . Note that this is a special case of  $A = \emptyset$  for  $\text{PAR}_n(A)$ .

## The Separators

Let us begin with a Chomsky normal form grammar  $\Gamma$ , with terminals  $T$  and productions

$$X_i \rightarrow Y_i Z_i, \quad i = 1, 2, \dots, n$$

in addition to certain productions of the form  $V \rightarrow a, a \in T$ . We construct a new grammar  $\Gamma_s$  which we call the *separator* of  $\Gamma$ . The terminals of  $\Gamma_s$  are the symbols of  $T$  together with  $2n$  new symbols  $(i, )_i, i = 1, 2, \dots, n$ . The productions of  $\Gamma_s$  are

$$X_i \rightarrow (i Y_i)_i Z_i, \quad i = 1, 2, \dots, n$$

as well as all of the productions in  $\Gamma$  of the form  $V \rightarrow a$  with  $a \in T$ .

## The Separators, Examples

As an example, let  $\Gamma$  have the productions

$$\begin{aligned} S &\rightarrow XY, & S &\rightarrow YX, & Y &\rightarrow ZZ, \\ X &\rightarrow a, & Z &\rightarrow a. \end{aligned}$$

The productions of  $\Gamma_s$  can be written as

$$\begin{aligned} S &\rightarrow (X)Y, & S &\rightarrow [Y]X, & Y &\rightarrow \{Z\}Z, \\ X &\rightarrow a, & Z &\rightarrow a. \end{aligned}$$

where we use  $(, ), [, ],$  and  $\{, \}$  in place for the numbered brackets.

## Ambiguity in Context-free Grammars

**Definition.** A context-free grammar  $\Gamma$  is called *ambiguous* if there is a word  $u \in L(\Gamma)$  that has two different leftmost derivations in  $\Gamma$ . If  $\Gamma$  is not ambiguous, it is said to be *unambiguous*.  $\square$  Note that grammar  $\Gamma$  in the last slide is ambiguous: There are two leftmost derivations for  $aaa$ :

$$S \Rightarrow XY \Rightarrow aY \Rightarrow aZZ \Rightarrow aaZ \Rightarrow aaa$$

$$S \Rightarrow YX \Rightarrow ZZX \Rightarrow aZX \Rightarrow aaX \Rightarrow aaa$$

However, for grammar  $\Gamma_s$ , the two derivations become

$$S \Rightarrow (X)Y \Rightarrow (a)Y \Rightarrow (a)\{Z\}Z \Rightarrow (a)\{a\}Z \Rightarrow (a)\{a\}a$$

$$S \Rightarrow [Y]X \Rightarrow [\{Z\}Z]X \Rightarrow [\{a\}Z]X \Rightarrow [\{a\}a]X \Rightarrow [\{a\}a]a$$

That is,  $\Gamma_s$  *separates* the two derivations in  $\Gamma$ . The bracketing in the words  $(a)\{a\}a$  and  $[\{a\}a]a$  enables their respective derivation trees to be recovered.

## Separated then Erased

If we write  $P$  for the set of brackets  $(, )_i, i = 1, 2, \dots, n$ , then clearly  $\Gamma = \text{Er}_P(\Gamma_s)$ . Hence, by Theorem 5.5, we conclude immediately that **Theorem 7.2.**  $\text{Er}_P(L(\Gamma_s)) = L(\Gamma)$ .  $\square$  In addition, we can also prove the following four lemmas about some relationship between languages  $L(\Gamma_s)$  and  $\text{PAR}_n(T)$ .

### Lemma 1

**Lemma 1.**  $L(\Gamma_s) \subseteq \text{PAR}_n(T)$ . *Proof.* We want to show that if  $X \Rightarrow_{\Gamma_s}^* w \in (T \cup P)^*$  for any variable  $X$ , the  $w \in \text{PAR}_n(T)$ . The proof is by an induction on the length of a derivation of  $w$  from  $X$  in  $\Gamma_s$ . If the length is 2, then  $w$  is a single terminal and the result is clear. Otherwise, we write

$$X = X_1 \Rightarrow_{\Gamma_s} ({}_i Y_i) {}_i Z_i \Rightarrow_{\Gamma_s}^* ({}_i u) {}_i v = w,$$

where  $Y_i \Rightarrow_{\Gamma_s}^* u$  and  $Z_i \Rightarrow_{\Gamma_s}^* v$ . By the induction hypothesis,  $u, v \in \text{PAR}_n(T)$ . By b and c of Theorem 7.1, so is  $w$ .  $\square$  To proceed further, we need to define a new context-free grammar  $\Delta$ , which is related to  $\Gamma_s$ .