Context-Free languages (part IV)

Prof. Dan A. Simovici

UMB







Other Closure Properties of L<sub>2</sub>

The main result of this section is a proof that the class  ${\cal R}$  of regular languages coincides with  ${\cal L}_3.$ 

### Theorem

Let G be a type-3 grammar, and let L be the language generated by G. There is a transition system T such that L = L(T).

## Proof

Suppose that  $G = (A_N, A_T, S, P)$  is a type-3 grammar. Define the transition system  $\mathfrak{T} = (A_T, A_N \cup \{Z\}, \theta, S, \{Z\})$ , where Z is a new symbol,  $Z \notin A_N \cup A_T$ , and

$$\theta = \{(X, u, Y) \mid X \to uY \in P\} \\ \cup \{(X, u, Z) \mid X \to u \in P\}.$$

Let  $w \in L(G)$ . There exists a derivation

$$S \underset{G}{\Rightarrow} u_0 X_{i_0} \underset{G}{\Rightarrow} u_0 u_1 X_{i_1} \cdots \underset{G}{\Rightarrow} u_0 u_1 \cdots u_{n-1} X_{i_{n-1}} \underset{G}{\Rightarrow} u_0 u_1 \cdots u_{n-1} u_n,$$

where  $w = u_0 \cdots u_{n-1}u_n$ . The productions used in this derivation are  $S \to u_0 X_{i_0}$ ,  $X_{i_{p-1}} \to u_p X_{i_p}$  for  $1 \leq p \leq n-1$ , and  $X_{i_{n-1}} \to u_n$ . Therefore, the triples

$$(S, u_0, X_{i_0}), (X_{i_0}, u_1, X_{i_1}), \dots, (X_{i_{n-2}}, u_{n-1}, X_{i_{n-1}}), (X_{i_{n-1}}, u_n, Z)$$

must all be in  $\theta$ , which implies that  $(S, u_0 \cdots u_n, Z) \in \theta^*$ . Since Z is a final state of T, we have  $u \in L(T)$ , so  $L(G) \subseteq L(T)$ .

Conversely, if  $u \in L(\mathcal{T})$ , then  $(S, u, Z) \in \theta^*$ . Taking into account the definition of  $\theta$ , there are *n* intermediate states in  $\mathcal{T}, X_{i_0}, \ldots, X_{i_{n-1}}$  such that  $u = u_0 \cdots u_n$  and the triples

$$(S, u_0, X_{i_0}), (X_{i_0}, u_1, X_{i_1}), \dots, (X_{i_{n-2}}, u_{n-1}, X_{i_{n-1}}), (X_{i_{n-1}}, u_n, Z)$$

exist in  $\theta$ . This implies the existence in P of the productions

$$S \to u_0 X_{i_0}, X_{i_0} \to u_1 X_{i_1}, \dots, X_{i_{n-2}} \to u_{n-1} X_{i_{n-1}}, X_{i_{n-1}} \to u_n$$

Using these productions we obtain the derivation

$$S \underset{G}{\Rightarrow} u_0 X_{i_0} \underset{G}{\Rightarrow} u_0 u_1 X_{i_1} \cdots \underset{G}{\Rightarrow} u_0 u_1 \cdots u_{n-1} X_{i_{n-1}} \underset{G}{\Rightarrow} u_0 u_1 \cdots u_{n-1} u_n,$$

which implies that  $x \in L(\mathcal{T})$ . This proves the inclusion  $L(\mathcal{T}) \subseteq L(G)$ .

### Theorem

For every regular language L there is a type-3 grammar G such that L(G) = L.

#### Proof.

Let  $\mathcal{M} = (A, Q, \delta, q_0, F)$  be a dfa such that  $L = L(\mathcal{M})$ . The type-3 grammar  $G = (Q, A, q_0, P)$  whose productions are

$$egin{array}{ll} q 
ightarrow aq' & ext{for each } q,q',a ext{ with } q' = \delta(q,a) \ q 
ightarrow \lambda & ext{for each } q \in F. \end{array}$$

generates  $L(\mathcal{M})$ .

### Corollary

The class  $\mathcal{L}_3$  coincides with the class  $\mathfrak{R}$  of regular languages.

Recall the Pumping Lemma for context-free languages:

#### Theorem

Let L be a context-free language. There exists a number  $n_L \in \mathbb{N}$  such that if  $w \in L$  and  $|w| \ge n_L$ , then we can write

w = xyzut

such that  $|y| \ge 1$  or  $|u| \ge 1$ ,  $|yzu| \le n_L$  and  $xy^n zu^n t \in L$  for all  $n \in \mathbb{N}$ .

This is a necessary condition for the "context-freeness" of a language.

## The Special Case of One-symbol Alphabets

Let  $A = \{a\}$  be an one-symbol alphabet.

- Word concatenation in A\* is commutative.
- The formulation of the Pumping Lemma in this special case:
   Let L be a context-free language. There exists a number n<sub>L</sub> ∈ N such that if w ∈ L and |w| ≥ n<sub>L</sub>, then we can write

$$w = rs$$

such that  $1 \leq |s| \leq n_G$  and  $rs^n \in L(G)$  for all  $n \in \mathbb{N}$ . Note that  $r \in L$  (since we can take n = 0). If  $|r| > n_L$  the same pumping lemma can be applied to r, and  $r = r_1w_1$  with  $|w_1| \le n_L$  such  $r_1w_1^{n_1} \in L$  for  $n_1 \in \mathbb{N}$ . Again  $r_1 \in L$  (for n = 0), etc. This leads to a stronger form of the Pumping Lemma for languages over one-symbol alphabets.

If *L* is a context-free language on an one-symbol alphabet, there exists a number  $n_L$  such that every word  $w \in L$  with  $|w| \ge n_L$  can be written as

$$w = rs_1s_2\cdots s_k,$$

where  $|r|, |s_1|, \ldots, |s_k| \leqslant n_L$  and

$$rs_1^{n_1}\cdots s_k^{n_k}\in L$$

for  $n_1, \ldots, n_k \in \mathbb{N}$ .

Note that the set  $K_n(L)$  of words in L shorter than  $n_L$  is finite, so it is a regular language. Since  $L = (L \cap K_n(L)) \cup (L - K_n(L))$ , and the set  $L - K_n(L)$  has the form  $\{w_1, w_2, \ldots, w_n\}^*$ , where  $w_1, \ldots, w_n$  are the words that can be "pumped", it follows that L is a regular language.

#### Theorem

Let  $s : A^* \longrightarrow B^*$  be a substitution. If s(a) is a context-free language for every  $a \in A$  and  $L \subseteq A^*$  is a context-free language, then s(L) is a context-free language.

## Proof

Suppose that L = L(G), where  $G = (A_N, A, S, P)$  is a context-free grammar and let s(a) is generated by the context-free grammar  $G_a = (A_N^a, B, S_a, P_a)$  for  $a \in A$ .

We may assume that the sets of nonterminal symbols  $A_N^a$  are pairwise disjoint.

Let P' be the set of productions obtained from P as follows. In each production of P replace every letter  $a \in A$  by the nonterminal  $S_a$ . We claim that the language s(L) is generated by the grammar  $G' = (A_N \cup \bigcup_{a \in A} A_N^a, B, S, P' \cup \bigcup_{a \in A} P_a).$ 

Let  $y \in s(L)$ . There exists a word  $x = a_{i_0} \dots a_{i_{n-1}} \in L$  such that  $y \in s(x)$ . This means that  $y = y_0 \dots y_{n-1}$ , where  $y_k \in s(a_{i_k}) = L(G_{a_{i_k}})$  for  $0 \leq k \leq n-1$ . Thus, we have the derivations  $S_{a_{i_k}} \stackrel{*}{\underset{G_{a_{i_k}}}{\longrightarrow}} y_k$  for  $0 \leq k \leq n-1$ , and the same derivations can be done in G'. Consequently, we obtain the derivation

$$S \stackrel{*}{\Rightarrow}_{G'} S_{a_{i_0}} \dots S_{a_{i_{n-1}}} \stackrel{*}{\Rightarrow}_{G'} y_0 \dots y_{n-1} = y,$$

which implies  $y \in L(G')$ , so  $s(L) \subseteq L(G')$ .

Conversely, if  $y \in L(G')$ , then any derivation  $S \stackrel{*}{\underset{G'}{\Rightarrow}} y$  is of the previous form.

The word y can be written as  $y = y_0 \dots y_{n-1}$ , where  $S_{a_{i_k}} \stackrel{*}{\Longrightarrow} g_k$  for  $0 \leq k \leq n-1$ , so  $y_k \in L(G_{a_{i_k}}) = s(a_{i_k})$  for  $0 \leq k \leq n-1$ . This implies  $y = y_0 \dots y_{n-1} \in s(a_{i_0} \dots s(a_{i_{n-1}}) = s(x) \in s(L)$ , so  $L(G') \subseteq s(L)$ . Since s(L) = L(G'), it follows that s(L) is a context-free language.

#### Corollary

If  $h : A^* \longrightarrow B^*$  is a morphism and  $L \subseteq A^*$  is a context-free language, then h(L) is a context-free language.

The class  $\mathcal{L}_2$  is closed with respect to inverse morphic images. In other words, if  $h: B^* \longrightarrow A^*$  is a morphism, and  $L \subseteq A^*$  is a context-free language, then  $h^{-1}(L)$  is a context-free language.

### Proof

Suppose that  $B = \{b_0, \ldots, b_{m-1}\}$  and that  $h(b_i) = x_i$  for  $0 \le i \le m-1$ . Let  $B' = \{b'_0, \ldots, b'_{m-1}\}$ , and let *s* be the substitution given by  $s(a) = B'^* a B'^*$  for  $a \in A$ .



$$B = \{b_0, \dots, b_{m-1}\}$$

$$(c \cup B)'^* \xrightarrow{h_2} B^* \xrightarrow{h(b_i) = x_i} A^*$$

$$h_1 \qquad g \qquad for all index i \leq m\}$$

$$H = \{b_i x_i | 0 \leq i \leq m\}$$

$$B' = \{b'_0, \dots, b'_{m-1}\}$$

Consider the finite language  $H = \{b'_i x_i \mid 0 \le i \le m-1\}$  in  $(B' \cup A)^*$  and the mapping  $g : \mathcal{P}(A^*) \longrightarrow \mathcal{P}((A \cup B')^*)$  given by  $g(L) = s(L) \cap H^*$ . Define  $h_1 : (A \cup B')^* \longrightarrow (\{c\} \cup B)^*$  and  $h_2 : (\{c\} \cup B)^* \longrightarrow B^*$  by  $h_1(a) = c$  for  $a \in A$ ,  $h_1(b') = b$  for all  $b' \in B'$ , and  $h_2(c) = \lambda$ ,  $h_2(b) = b$  for  $b \in B$ . We claim that for every language  $L \in \mathcal{P}(A)$  such that  $\lambda \notin L$ ,  $h^{-1}(L) = h_2(h_1(g(L)))$  and hence,  $h^{-1}(L)$  is context-free. This follows from the following equivalent statements:

• 
$$u = b_{i_0} \cdots b_{i_{k-1}} \in h^{-1}(L);$$
  
•  $h(u) = x_{i_0} \cdots x_{i_{k-1}} \in L;$   
•  $b'_{i_0} x_{i_0} \cdots b'_{i_{k-1}} x_{i_{k-1}} \in g(L);$   
•  $h_1(b'_{i_0} x_{i_0} \cdots b'_{i_{k-1}} x_{i_{k-1}}) = b_{i_0} c \cdots c \cdots b_{i_{k-1}} c \cdots c \in h_1(g(L));$   
•  $h_2(b_{i_0} c \cdots c \cdots b_{i_{k-1}} c \cdots c) = b_{i_0} \cdots b_{i_{k-1}} = u \in h_2(h_1(g(L))).$ 

If  $\lambda \in L$ , the language  $L - \{\lambda\}$  is context-free, so  $h^{-1}(L - \{\lambda\})$  is also context-free. Note that  $h^{-1}(L) = h^{-1}(L - \{\lambda\}) \cup h^{-1}(\{\lambda\})$  and that  $h^{-1}(\{\lambda\}) = \{a \in A \mid h(a) = \lambda\}^*$ . Since  $h^{-1}(\{\lambda\})$  is regular it follows that  $h^{-1}(L)$  is context-free.

## Reminder

We defined the shuffle of languages

### Definition

Let A be an alphabet and let G, K be two languages over A. The *shuffle* of G and K is the language

shuffle
$$(G, K) = \{x_0y_0x_1y_1\cdots x_{n-1}y_{n-1} \mid x_0x_1\cdots x_{n-1} \in G \\ and y_0y_1\cdots y_{n-1} \in K\}.$$

We proved

#### Theorem

There is an alphabet B and there exist three morphisms g, k, h from B<sup>\*</sup> to A<sup>\*</sup> such that h is a very fine morphism, g, k are fine morphisms and shuffle(G, K) =  $h(g^{-1}(G) \cap k^{-1}(K))$ .

#### Corollary

Let  $L \subseteq A^*$  be a context-free language and let  $R \subseteq A^*$  be a regular language. Then, shuffle(L, R) is a context-free language.