

## A Language Equation and Its Applications

京都産業大学・理学部 伊藤 正美

Masami Ito

Department of Mathematics

Kyoto Sangyo University

Kyoto 603-8555, Japan

Email: ito@ksuvsx0.kyoto-su.ac.jp

Let  $u, v \in X^*$  be words over an alphabet  $X$ . Then the set  $\{u_1v_1u_2v_2 \dots u_nv_n \mid u = u_1u_2 \dots u_n, v = v_1v_2 \dots v_n, u_1, v_1, u_2, v_2, \dots, u_n, v_n \in X^*, n \geq 1\}$  is called the *shuffle product* of  $u$  and  $v$ , and denoted by  $u \diamond v$ . For languages  $A, B \subseteq X^*$ , the set  $A \diamond B = \bigcup_{u \in A, v \in B} u \diamond v$  is called the *shuffle product* of  $A$  and  $B$ . In this paper, we consider the following problem: Let  $A, B \subseteq X^*$  be regular languages. Then can we obtain a solution  $C \subseteq X^*$  of the language equation  $A = B \diamond C$ ? Obviously, this problem is equivalent to the shuffle decomposition problem for regular languages. Regarding definitions and notations concerning formal languages and automata, not defined in this paper, refer, for instance, to [1].

Now let  $\mathcal{A} = (S, X, \delta, s_0, F)$  be a finite automaton with  $\mathcal{T}(\mathcal{A}) = A$  and let  $\mathcal{B} = (T, X, \gamma, t_0, G)$  be a finite automaton with  $\mathcal{T}(\mathcal{B}) = B$ . We will look for a regular language  $C$  over  $X$  such that  $A = B \diamond C$ . By  $\overline{X}$ , we denote the language  $\{\bar{a} \mid a \in X\}$  with  $X \cap \overline{X} = \emptyset$ . Let  $\overline{\mathcal{B}} = (T, X \cup \overline{X} \cup \{\#\}, \overline{\gamma}, t_0, G)$  where  $\overline{\gamma}$  is defined as follows:

For  $t \in T$  and  $a \in X$ ,  $\overline{\gamma}(t, a) = t$ ,  $\overline{\gamma}(t, \bar{a}) = \gamma(t, a)$ . Moreover,  $\overline{\gamma}(t, \#) = t$  if  $t \in G$ .

Then the following can be easily shown.

**Fact 1** Let  $a_1a_2 \dots a_n \in X^*$  where  $a_i \in X, i = 1, 2, \dots, n$ . Then  $a_1a_2 \dots a_n \in \mathcal{T}(\mathcal{B})$  if and only if  $u_1\bar{a}_1u_2\bar{a}_2 \dots u_n\bar{a}_nu_{n+1}\# \in \mathcal{T}(\overline{\mathcal{B}})$  where  $u_1, u_2, \dots, u_n \in X^*$ .

Let  $\mathcal{A}_1 = (\overline{S}, X \cup \overline{X} \cup \{\#\}, \overline{\delta}, s_0, \{\alpha, \omega\})$  and let  $\mathcal{A}_2 = (\overline{S}, X \cup \overline{X} \cup \{\#\}, \overline{\delta}, s_0, \{\alpha\})$  where  $\overline{S} = (\bigcup_{a \in X \cup \{\epsilon\}} S^{(a)}) \cup \{\alpha, \omega\}$ . Here  $S^{(\epsilon)}$  is regarded as  $S$  where  $\epsilon$  is the empty word. For  $s \in S, t \in S \setminus F, t' \in F, a \in X \cup \{\epsilon\}, b \in X$  and  $\{\#\}$ ,  $\overline{\delta}$  is defined as follows:

$\bar{\delta}(s^{(a)}, b) = \delta(s, b)^{(a)}$ ,  $\bar{\delta}(s^{(a)}, \bar{b}) = \delta(s, b)^{(b)}$ ,  $\bar{\delta}(t^{(a)}, \#) = \{\alpha\}$  and  $\bar{\delta}(t'^{(a)}, \#) = \{\omega\}$ .

We consider the following two automata:

$\mathcal{C}_1 = (\bar{S} \times T, X \cup \bar{X} \cup \{\#\}, \bar{\delta} \times \bar{\gamma}, (s_0, t_0), \{\alpha, \omega\} \times G)$ ,  $\mathcal{C}_2 = (\bar{S} \times T, X \cup \bar{X} \cup \{\#\}, \bar{\delta} \times \bar{\gamma}, (s_0, t_0), \{\alpha\} \times G)$  where  $\bar{\delta} \times \bar{\gamma}((\bar{s}, t), a) = (\bar{\delta}(\bar{s}, a), \bar{\gamma}(t, a))$  for  $(\bar{s}, t) \in \bar{S} \times T$  and  $a \in X$ .

Now consider the following homomorphism  $\rho$  of  $(X \cup \bar{X} \cup \{\#\})^*$  into  $X^*$ :

$\rho(a) = a$  for  $a \in X$ ,  $\rho(\bar{a}) = \epsilon$  for  $a \in X$  and  $\rho(\#) = \epsilon$ .

**Lemma 1** *Automata accepting the languages  $\rho(\mathcal{T}(\mathcal{C}_1))$  and  $\rho(\mathcal{T}(\mathcal{C}_2))$  can be effectively constructed.*

*Proof* Let  $i = 1, 2$ . From  $\mathcal{C}_i$ , we can construct a regular grammar  $\mathcal{G}_i$  such that  $\mathcal{L}(\mathcal{G}_i) = \mathcal{T}(\mathcal{C}_i)$  with the production rules of the form  $A \rightarrow aB$  ( $A, B$  are variables and  $a \in X \cup \bar{X} \cup \{\#\}$ ). Replacing every rule of the form  $A \rightarrow aB$  in  $\mathcal{G}_i$  by  $A \rightarrow \rho(a)B$ , we can obtain a new grammar  $\mathcal{G}'_i$ . Then it is clear that  $\rho(\mathcal{L}(\mathcal{C}_i)) = \mathcal{L}(\mathcal{G}'_i)$ . Using this grammar  $\mathcal{G}'_i$ , we can construct an automaton  $\mathcal{D}_i$  such that  $\mathcal{T}(\mathcal{D}_i) = \mathcal{T}(\mathcal{G}'_i)$  i.e.  $\rho(\mathcal{T}(\mathcal{C}_i)) = \mathcal{T}(\mathcal{D}_i)$ . Notice that all the above procedures are effectively done. This completes the proof of the lemma.

Let  $B, C \subseteq X^*$ . By  $B \diamond C$  we denote the shuffle product of  $B$  and  $C$ , i.e.  $\{u_1v_1u_2v_2 \dots u_nv_n \mid u = u_1u_2 \dots u_n \in B, v = v_1v_2 \dots v_n \in C\}$ .

**Proposition 1** *Let  $u \in X^*$ . Then  $\{u\} \diamond B \subseteq A$  if and only if  $u \in \rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2))$ .*

*Proof* ( $\Rightarrow$ ) Let  $u = u_1u_2 \dots u_nu_{n+1} \in X^*$  and let  $a_1a_2 \dots a_n \in B$  where  $u_1, u_2, \dots, u_n, u_{n+1} \in X^*$  and  $a_1, a_2, \dots, a_n \in X$ . Then  $\bar{\delta} \times \bar{\gamma}((s_0, t_0), u_1\bar{a}_1u_2\bar{a}_2 \dots u_n\bar{a}_nu_{n+1}\#) = (\bar{\delta}(s_0, u_1\bar{a}_1u_2\bar{a}_2 \dots u_n\bar{a}_nu_{n+1}\#), \bar{\gamma}(t_0, u_1\bar{a}_1u_2\bar{a}_2 \dots u_n\bar{a}_nu_{n+1}\#)) = (\bar{\delta}(\delta(s_0, u_1a_1u_2a_2 \dots u_na_nu_{n+1})^{(a_n)}, \#), \bar{\gamma}(\gamma(s_0, a_1a_2 \dots a_n)^{(a_n)}, \#)) = (\omega, \gamma(t_0, a_1a_2 \dots a_n)) \in \{\omega\} \times G$ . Therefore,  $u_1\bar{a}_1u_2\bar{a}_2 \dots u_n\bar{a}_nu_{n+1}\# \in \mathcal{T}(\mathcal{C}_1) \setminus \mathcal{T}(\mathcal{C}_2)$ . Hence  $u = u_1u_2 \dots u_nu_{n+1} = \rho(u_1\bar{a}_1u_2\bar{a}_2 \dots u_n\bar{a}_nu_{n+1}\#) \in \rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2))$ .

( $\Leftarrow$ ) Suppose that  $u \diamond B \subseteq A$  does not hold though  $u \in \rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2))$ . Then there exist  $u = u_1u_2 \dots u_nu_{n+1} \in X^*$  and  $a_1a_2 \dots a_n \in B$  such that  $u_1a_1u_2a_2 \dots u_na_nu_{n+1} \notin A$ . Hence  $\bar{\gamma}(t_0, u_1\bar{a}_1u_2\bar{a}_2 \dots u_n\bar{a}_nu_{n+1}\#) = \bar{\gamma}(\gamma(t_0, a_1a_2 \dots a_n), \#) = \gamma(t_0, a_1a_2 \dots a_n) \in G$ . On the other hand, since  $u_1a_1u_2a_2 \dots u_na_nu_{n+1} \notin A$ , we have  $\bar{\delta}(s_0, u_1\bar{a}_1u_2\bar{a}_2 \dots u_n\bar{a}_nu_{n+1}\#) = \bar{\delta}(\delta(s_0, u_1a_1u_2a_2 \dots u_na_nu_{n+1})^{(a_n)}, \#) = \{\alpha\}$ .

$u_{n+1})^{(a_n)}, \#) = \{\alpha\}$ . Hence  $\bar{\delta} \times \bar{\gamma}((s_0, t_0), u_1 \bar{a}_1 u_2 \bar{a}_2 \dots u_n \bar{a}_n u_{n+1} \#) \in \{\alpha\} \times G$ , i.e.  $u_1 \bar{a}_1 u_2 \bar{a}_2 \dots u_n \bar{a}_n u_{n+1} \# \in \mathcal{T}(\mathcal{C}_2)$ . Therefore,  $u = \rho(u_1 \bar{a}_1 u_2 \bar{a}_2 \dots u_n \bar{a}_n u_{n+1} \#) \in \rho(\mathcal{T}(\mathcal{C}_2))$ . On the other hand, it is obvious that  $u_1 \bar{a}_1 u_2 \bar{a}_2 \dots u_n \bar{a}_n u_{n+1} \# \in \mathcal{T}(\mathcal{C}_1)$ . Thus  $u \notin \rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2))$ , a contradiction. Consequently, the proposition must hold true.

**Corollary** *In the above,  $B \diamond (\rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2))) \subseteq A$ .*

Let  $L \subseteq X^*$  be a regular language over  $X$ . By  $\#L$ , we denote the number  $\min\{|S| \mid \exists \mathcal{A} = (S, X, \delta, s_0, F), L = \mathcal{T}(\mathcal{A})\}$  where  $|S|$  denotes the cardinality of  $S$ . Moreover,  $\mathcal{I}(n, X)$  denotes the class of languages  $\{L \subseteq X^* \mid \#L \leq n\}$ .

**Theorem 1** *Let  $A \subseteq X^*$  and let  $n$  be a positive integer. Then it is decidable whether there exist nontrivial regular languages  $B \in \mathcal{I}(n, X)$  and  $C \subseteq X^*$  such that  $A = B \diamond C$ . Here a language  $D \subseteq X^*$  is said to be nontrivial if  $D \neq \{\epsilon\}$ .*

*Proof* Let  $A \subseteq X^*$  be a regular language. Assume that there exist nontrivial regular languages  $B \in \mathcal{I}(n, X)$  and  $C \subseteq X^*$  such that  $A = B \diamond C$ . Then, by Proposition 1 and its corollary,  $C \subseteq \rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2))$  and  $B \diamond (\rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2))) \subseteq A$ . Hence  $A = B \diamond (\rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2)))$ . Thus we have the following algorithm: (1) Choose a nontrivial regular language  $B \subseteq X^*$  from  $\mathcal{I}(n, X)$  and construct the language  $\rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2))$  (see Lemma 1). (2) Let  $C = \rho(\mathcal{T}(\mathcal{C}_1)) \setminus \rho(\mathcal{T}(\mathcal{C}_2))$ . (3) Compute  $B \diamond C$ . (4) If  $A = B \diamond C$ , then the output is "YES" and "NO", otherwise. (4) If the output is "NO", then choose another element in  $\mathcal{I}(n, X)$  as  $B$  and continue the procedures (1) - (3). (5) Since  $\mathcal{I}(n, X)$  is a finite set, the above process terminates after a finite-step trial. Once one gets the output "YES", then there exist nontrivial regular languages  $B \in \mathcal{I}(n, X)$  and  $C \subseteq X^*$  such that  $A = B \diamond C$ . Otherwise, there are no such languages.

Let  $n$  be a positive integer. By  $\mathcal{F}(n, X)$ , we denote the class of finite languages  $\{L \subseteq X^* \mid \max\{|u| \mid u \in L\} \leq n\}$  where  $|u|$  is the length of  $u$ . Then the following result by C. Cămpăanu et al. ([2]) can be obtained as a corollary of the above theorem.

**Corollary** *For a given positive integer  $n$  and a regular language  $A \subseteq X^*$ , the problem whether  $A = B \diamond C$  for a nontrivial language  $B \in \mathcal{F}(n, X)$  and a nontrivial regular language  $C \subseteq X^*$  is decidable.*

*Proof* Obvious from the fact that  $\mathcal{F}(n, X) \subseteq \mathcal{I}(|X|^{n+1}, X)$ .

## References

- [1] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, Reading MA, 1979.
- [2] C. C ampeanu, K. Salomaa, S. V agv olgyi, Shuffle quotient and decompositions, Lecture Notes in Computer Science (Springer), to appear.