Quotients of Context-Free Languages*

SEYMOUR GINSBURG

System Development Corporation, Santa Monica, California

AND

EDWIN H. SPANIER

University of California, Los Angeles

Abstract. The following results on the quotient of context-free languages (CFL) are shown: (1) It is recursively unsolvable to determine for arbitrary CFL whether the quotient of one by another is a CFL. (2) If either set is regular and the other is a CFL, then the quotient is a CFL.

1. *Introduction*

Among the operations under investigation by the SHARE Theory of Information Handling Committee is that of quotient. This paper sets forth some results about quotients of context-free languages (abbreviated CFL), i.e., quotients of components of ALGOL-like languages. These results, proved in Section 3, are the following:

(1.1) It is recursively unsolvable to determine for arbitrary CFL whether the quotient of one by another is again a CFL.

(1.2) If either set is regular and the other is a CFL, then the quotient is a CFL.

2. *Preliminaries*

Let Σ be a finite nonempty set, or alphabet, and let $\theta(\Sigma)$ be the free semigroup with identity ϵ generated by Σ. (Thus $\theta(\Sigma)$ is the set of all finite sequences, or words, of Σ and ϵ is the empty sequence.) We shall be considering subsets of $\theta(\Sigma)$. If A and B are subsets of $\theta(\Sigma)$, then so is the *product* $AB = \{ab/a \text{ in } A, b \text{ in } B\}$.

A *grammar* G is a 4-tuple (V, P, Σ, S), where V is a finite set, Σ is a nonempty subset of V, S is an element of V-Σ, and P is a finite set of ordered pairs of the form (ξ, w) with ξ in V-Σ and w in $\theta(V)$. P is called the set of *production* of G. An element (ξ, w) in P is denoted by $\xi \rightarrow w$. If x and y are in $\theta(V)$, then we write $x \Rightarrow y$ if either $x = y$ or there exists a sequence $x = x_1, x_2, \cdots, x_n = y$ $(n > 1)$ of elements in $\theta(V)$ with the following property: For each $i < n$ there exists a_i, b_i, ξ_i, w_i such that $x_i = a_i\xi_ib_i$, $x_{i+1} = a_iw_ib_i$ and $\xi_i \rightarrow w_i$. The *language* generated by G, denoted by $L(G)$, is the set of words $\{w/S \Rightarrow w, w \text{ in } \theta(\Sigma)\}$. A

* This work was sponsored in part by the Air Force Cambridge Research Laboratories, Office of Aerospace Research, under Contract AF 19(628)-485.

context-free language (over Σ) is a language $L(G)$ generated by some grammar $G = (V, P, \Sigma, S)$.

The concept of CFL was introduced by Chomsky [2] in his study of natural languages. It has since been shown that context-free languages are identical with the components in the "ALGOL-like" artificial languages which arise in data processing [5]. As such, their properties are currently being studied [6, 7, 11, 12].

A special kind of context-free language called a regular set has been introduced [8] in connection with the theory of automata. We now present the relevant definitions of these concepts. An *automaton* [10] is a 5-tuple $A = (K, \Sigma, \delta, s_0, F)$, where

(i) K is a finite nonempty set (called the set of *states*);
(ii) Σ is a finite nonempty set (called the set of *inputs*);
(iii) δ is a mapping from $K \times \Sigma$ into K (called the *next state function*);
(iv) s_0 is an element of K (called the *start* state);
(v) F is a subset of K (called the set of *final* states).

Given such an automaton the next state function δ can be extended to a mapping, also denoted by δ, from $K \times \theta(\Sigma)$ to K, inductively by

$$\delta(q, \epsilon) = q \quad \text{for} \quad q \text{ in } K$$

and

$$\delta(q, I_1 I_2 \cdots I_k) = \delta(\delta(q, I_1 I_2 \cdots I_{k-1}), I_k) \quad \text{for} \quad q \text{ in } K,\ I_i \text{ in } \Sigma,\ k \geqq 2.$$

For an automaton A denote by $T(A)$ the set $\{w/w \text{ in } \theta(\Sigma),\ \delta(s_0, w) \text{ in } F\}$. A subset $R \subseteq \theta(\Sigma)$ is said to be *regular* (or Σ-*regular* when there is a need to distinguish Σ) if there is an automaton $A = (K, \Sigma, \delta, s_0, F)$ such that $R = T(A)$.

It is known [3] that every regular set is a CFL. Since a regular set is a language generated by a finite state device, it is sometimes called a *finite state language.*

The concept of quotient mentioned in the introduction is now defined. If X and Y are subsets of $\theta(\Sigma)$, then the *right quotient* of X and Y, denoted by X/Y, is the subset of $\theta(\Sigma)$ defined by $X/Y = \{w/wy \text{ in } X \text{ for some } y \text{ in } Y\}$. Similarly the *left quotient* $Y \backslash X = \{w/yw \text{ in } X \text{ for some } y \text{ in } Y\}$. We shall be concerned with the right quotient, but all the results have obvious analogues for the left quotient. The following elementary properties are easily verified using the definitions.

(2.1) $X/(Y \cup Z) = X/Y \cup X/Z$.
(2.2) $(X \cup Z)/Y = X/Y \cup Z/Y$.
(2.3) $X/YZ = (X/Z)/Y$.
(2.4) $(XZ)/Y = X(Z/Y) \cup X/(Y/Z)$.

We are interested in the question of whether or not the quotient of one CFL by another is a CFL and discuss this in the next section.

3. *Results*

We now show that it is recursively unsolvable to determine if the quotient of one CFL by another is a CFL. First, we treat the case where one of the CFL is a regular set.

It is noted without proof in [4] that if X and Y are both regular, then X/Y is also regular. We have the following extension of that result.

(3.1) THEOREM. *If X is regular and Y is arbitrary, then X/Y is regular.*

PROOF. If Y is empty, then X/Y is empty and thus regular. If Y is nonempty, let $X = T(A)$ where $A = (K, \Sigma, \delta, s_0, F)$. Let $F_0 = \{q/q \text{ in } K \text{ and } \delta(q, y) \text{ in } F \text{ for some } y \text{ in } Y\}$. It is readily seen that $X/Y = T(B)$, where $B = (K, \Sigma, \delta, s_0, F_0)$. Thus X/Y is regular.

Next consider the case where Y is regular and X is a CFL. First we establish a preliminary lemma which shows that any regular set can be defined by an automaton in which the start state is not the next state of any state.

(3.2) LEMMA. *If $A = (K, \Sigma, \delta, s_0, F)$ is an automaton, then there exists an automaton $A' = (K', \Sigma, \delta', s_0', F')$ such that $T(A) = T(A')$ and $\delta'(q, I) \neq s_0'$ for q in K' and I in Σ.*

PROOF. Let s_0' be an element not in K and let $K' = K \cup \{s_0'\}$. Define $F' \subseteq K'$ by

$$F' = \begin{cases} F \cup \{s_0'\} & \text{if } s_0 \text{ is in } F. \\ F & \text{if } s_0 \text{ is not in } F. \end{cases}$$

For I in Σ define $\delta'(s_0', I) = \delta(s_0, I)$ and $\delta'(q, I) = \delta(q, I)$ if q is in K. Clearly $A' = (K', \Sigma, \delta', s_0', F')$ has the desired properties.

(3.3) THEOREM. *If X is a CFL and Y is regular, then X/Y is a CFL.*

PROOF. If ϵ is in X, then $X = (X - \epsilon) \cup \epsilon$. Thus, by(2.2), $X/Y = (X - \epsilon)/Y \cup \epsilon/Y$. Now ϵ/Y is either empty or $\{\epsilon\}$. In either case it is a CFL. By [1, 5] it is known that $X - \epsilon$ is also a CFL. Since the finite union of CFL is again a CFL [1], it suffices to show that $(X - \epsilon)/Y$ is a CFL. Hence we need only prove the theory for the case where ϵ is not in X.

Let $A = (K, \Sigma, \delta, s_0, F)$ be an automaton such that $T(A) = Y$ and (by (3.2)) such that $\delta(q, I) \neq s_0$ for q in K, I in Σ. For each q in F let $T_q = \{w/\delta(s_0, w) = q, w \text{ in } \theta(\Sigma)\}$. Then Y is the finite union of the regular sets T_q and, by (2.1), $X/Y = \bigcup X/T_q$. Since a finite union of CFL is a CFL, it suffices to show that X/T_q is a CFL. Hence we need only prove the theorem for regular sets Y of the form $Y = T(A)$ where $A = (K, \Sigma, \delta, s_0, \{t\})$ (i.e., the set of final states of A consists of the single element t) and $\delta(q, I) \neq s_0$ for q in K, I in Σ.

If ϵ is not in X, then there exists a grammar $G = (V, P, \Sigma, S)$ such that $X = L(G)$ and P contains no production of the form $\xi \to \epsilon$ [1]. Let $Y = T(A)$ where $A = (K, \Sigma, \delta, s_0, \{t\})$ and $\delta(q, I) \neq s_0$ for q in K, I in Σ. Consider the grammar $G' = (V', P', \Sigma, S')$ where $V' = \Sigma \cup (K \times V \times K)$, $S' = (s_0, S, t)$, and P' consists of the following productions:

(1) $(s_0, x, s_0) \to x$ for each x in Σ.

(2) $(q, x, q') \to \epsilon$ if x is in Σ and $\delta(q, x) = q'$.

(3) $(q, x, q') \to (q, y_1, q_1)(q_1, y_2, q_2) \cdots (q_{n-1}, y_n, q')$ if $x \to y_1y_2 \cdots y_n$ is in P and $q_1, q_2, \cdots, q_{n-1}$ are in K.

We shall prove that $X/Y = L(G')$.

(a) To show that $L(G') \subseteq X/Y$ let w' be in $L(G')$. Then $(s_0, S, t) \Rightarrow w'$. Since a production of type (3) commutes with one of type (1) or (2), the sequence of productions yielding $(s_0, S, t) \Rightarrow w'$ can be arranged so that all the productions of type (3) precede those of types (1) and (2). Hence we may assume that by type (3) productions

$$(s_0, S, t) \Rightarrow (s_0, y_1, q_1)(q_1, y_2, q_2) \cdots (q_m, y_{m+1}, t)$$

and by types (1) and (2) productions

$$(s_0, y_1, q_1)(q_1, y_2, q_2) \cdots (q_m, y_{m+1}, t) \Rightarrow w'.$$

Since $(s_0, y_1, q_1)(q_1, y_2, q_2) \cdots (q_m, y_{m+1}, t) \Rightarrow w'$ by types (1) and (2), each y_i is in Σ. Since every type (3) production corresponds to a production of P, it follows that $S \Rightarrow y_1 \cdots y_{m+1}$ in G. Thus $y_1 \cdots y_{m+1}$ is in X. Furthermore, for each $1 \leqq i \leqq m+1$ either (q_{i-1}, y_i, q_i) is such that $q_{i-1} = q_1 = s_0$ or $\delta(q_{i-1}, y_i) = q_i$. Let j be the largest integer such that $q_j = s_0$. Because $\delta(q, I) \neq s_0$ for q in K, I in Σ, it follows that $\delta(q_i, y_{i+1}) = q_{i+1} \neq s_0$ for $i \geqq j$. Since w' is in $\theta(\Sigma)$, we see that $\delta(s_0, y_{j+1}y_{j+2} \cdots y_{m+1}) = t$. Thus $y_{j+1} \cdots y_{m+1}$ is in Y and $w' = y_1 \cdots y_j$. Since $w'y_{j+1} \cdots y_{m+1}$ is in X, w' is in X/Y.

(b) To show that $X/Y \subseteq L(G')$ let $x_1 \cdots x_m$ be an element of X/Y. Then there exists $y_1 \cdots y_n$ in Y such that $x_1 \cdots x_m y_1 \cdots y_n$ is in X. Since ϵ is not in X, either $x_1 \cdots x_m \neq \epsilon$ or $y_1 \cdots y_n \neq \epsilon$. First assume that neither is ϵ. Since $S \Rightarrow x_1 \cdots x_m y_1 \cdots y_n$ in G, we see that by type (3) productions we have

$$(q_0, S, t) \Rightarrow (q_0, x_1, q_0) \cdots (q_0, x_m, q_0)(q_0, y_1, q_1) \cdots q_{n-1}, y_n, t)$$

where q_i is defined to be $\delta(q_{i-1}, y_i)$ for $i \geqq 1$. Applying type (1) productions to (q_0, x_j, q_0) and type (2) productions to (q_{i-1}, y_i, q_i) we see that $S' \Rightarrow x_1 \cdots x_m$ in G'. Therefore $x_1 \cdots x_m$ is in $L(G')$. If $x_1 \cdots x_m = \epsilon$ (or $y_1 \cdots y_n = \epsilon$), then the above argument holds except that no productions of type (1) (or type (2)) need be applied to show that $S' \Rightarrow x_1 \cdots x_m$ in G'. In any case, $X/Y \subseteq L(G')$, which completes the proof.

(3.1) and (3.3) together establish (1.2). We now prove (1.1).

(3.4) THEOREM. *It is recursively unsolvable to determine for arbitrary CFL, X and Y, whether or not X/Y is a CFL.*

PROOF. Let $\Sigma = \{a, b, c\}$. For each positive integer n let $\bar{n} = ab^n$ (b^n is defined inductively by $b^1 = b$, $b^{j+1} = b^jb$ for $j \geqq 1$). For every n-tuple $w = (w_1, \cdots, w_n)$ of non-ϵ-words of $\theta(a, b)$ let

$$L(w) = \{cw_{i_1} \cdots w_{i_k}c\bar{i}_k \cdots \bar{i}_1 / k \geqq 1;\ 1 \leqq i_1, \cdots, i_k \leqq n\}.$$

Then $L(w)$ is a CFL. In fact, $L(w) = L(G)$ where $G = (\Sigma \cup \{\xi^{(1)}, \xi^{(2)}\}, P, \Sigma, \xi^{(2)})$ and P consists of the productions $\xi^{(1)} \to w_i\xi^{(1)}\bar{i}$, for $i \leqq i \leqq n$; $\xi^{(1)} \to w_ic\bar{i}$ for $i \leqq i \leqq n$; and $\xi^{(2)} \to c\xi^{(1)}$.

Let $y = (y_1, \cdots, y_n)$ and $z = (z_1, \cdots, z_n)$ be arbitrary n-tuples of non-ϵ-words of $\theta(a, b)$. It is obvious that $L(y)/L(z)$ either consists of ϵ or is empty according as there does or does not exist a sequence of integers $i_1, \cdots, i_k$ such that $y_{i_1} \cdots y_{i_k} = z_{i_1} \cdots z_{i_k}$. The existence of such a sequence of integers is the well-known Post Correspondence Problem and is recursively unsolvable [9].

Let L_1 and L_2 be arbitrary CFL. Then $L_1L(y)$ and $L_2L(z)$ are CFL since the product of CFL is a CFL [1]. It is easily seen (either directly from the definition or by applying (2.1), (2.3), (2.4)) that $L_1L(y)/L_2L(z)$ is L_1/L_2 or empty according to whether $L(y)/L(z)$ consists of ϵ or is empty. In particular, if L_1/L_2 is not a CFL, then $L_1L(y)/L_2L(z)$ is a CFL if and only if there does not exist a sequence of integers $i_1, \cdots, i_k$ such that $y_{i_1} \cdots y_{i_k} = z_{i_1} \cdots z_{i_k}$ and so is recursively unsolvable. Therefore, to complete the proof it suffices to exhibit particular CFL, L_1 and L_2, for which L_1/L_2 is not a CFL.

Consider the alphabet $\{a, b, c, d\}$. Let $1' = a$, $2' = b$, and $3' = c$. For all words x_1, x_2, x_3 in $\theta(a, b, c)$, let

$$L(x_1, x_2, x_3) = \{x_{i_1} \cdots x_{i_k} d i_k' \cdots i_1' / k \geqq 1; 1 \leqq i_1, \cdots, i_k \leqq 3\}.$$

Then $L(x_1, x_2, x_3)$ is a CFL. In fact, $L(x_1, x_2, x_3) = L(H)$, where $H = (\{a, b, c, d, \xi\}, P_H, \{a, b, c, d\}, \xi)$ and P_H consists of the productions

$$\xi \to x_i d i' \quad \text{and} \quad \xi \to x_i \xi i' \qquad \text{for } i = 1, 2, 3.$$

Therefore $L_1 = L(b^2, a^3, abc)$ and $L_2 = L(a, b, c)$ are CFL. We shall show that L_1/L_2 is not a CFL.

Let $Z = L_1/L_2$. Each word z in Z is obtained from words z_1 in L_1 and z_2 in L_2 satisfying $z_1 = zz_2$. Since each word in L_1 or L_2 contains the letter d exactly once, the terminal subwords starting from d in z_1 and in z_2 are the same. If the word da (or db) is a subword of z_1, then b^2da (or a^3db) occurs in z_1 and ada (or bdb) occurs in z_2. Either case contradicts the equation $z_1 = zz_2$. Thus the only letter which can occur immediately to the right of d in z_1 is c. Therefore z_1 must contain $abcdc$ as a subword and z_2 must contain cdc as a subword. Hence the shortest words which can occur as z_1, z_2 are $abcdc$ and cdc. Thus ab is in Z. In z_1 we see that cdc is preceded by b. Then any longer word for z_2 must contain $bcdcb$, and the corresponding z_1 must contain $a^3abcdcb$. Therefore a^4 is in Z. This line of reasoning can be continued inductively to provide a means of enumerating all the elements of Z. We find that Z consists of the sequence

$$ab,\ a^4,\ b^2a^3,\ b^4a^2,\ b^6a,\ b^8,\ a^3b^7,\ a^6b^6,\ \cdots,\ a^{24},\ \cdots$$

where to pass from one word x_i in the sequence to the next x_{i+1} we use the following rules:

(i) If $x_i = y_i a$, then $x_{i+1} = b^2 y_i$.

(ii) If $x_i = y_i b$, then $x_{i+1} = a^3 y_i$.

Thus a^n is in Z if and only if $n = 4.6^i$ for $i \geqq 0$. Let Z_0 be the set obtained by replacing each occurrence of a by a and each occurrence of b by the empty set. Then $Z_0 = \{a^n/n = 4 \cdot 6^i \text{ for } i \geqq 0\}$. By [1] it is known that if Z is a CFL then so is Z_0. But a set of the form $\{a^j/j \text{ in } A\}$ is a CFL if and only if A is ultimately periodic [5]. Since $\{4 \cdot 6^i/i \geqq 0\}$ is not ultimately periodic, Z_0 is not a CFL so neither is Z. Thus the theorem is proved.

In conclusion we state the following open problem:

A CFL is said to be sequential [5] if the elements of V-Σ may be labeled $x_1, \cdots, x_n$, with $x_n = S$, so that for each production $x_i \to ux_jv$ in P, $j \leqq i$. If X is a sequential CFL and Y is regular, is X/Y sequential?

REFERENCES

1. BAR-HILLEL, PERLES, AND SHAMIR. On formal properties of simple phrase structure grammars. *Zeit. Phonetik, Sprachwiss. Kommunikationsforsch. 14*, (1961), 143–172.
2. CHOMSKY, N. Three models for the description of language. *IRE Trans. IT2* (1956), 113–124.
3. ——. On certain formal properties of grammars. *Inform. Contr. 2* (1959), 137–167.
4. ELGOT AND RUTLEDGE, Operations on finite automata. Proc. Second Ann. Symp. Switching Circuit Theory and Logical Design, Detroit, Oct. 1961, 129–132.
5. GINSBURG, S., AND RICE, H. G. Two families of languages related to ALGOL. *J. ACM 9* (1962), 350–371.
6. GINSBURG, S., AND ROSE, G. F. Operations which preserve definability in languages. *J. ACM 10* (1963), 175–195.
7. ——. Some recursively unsolvable problems in ALGOL-like languages. *J. ACM 10* (1963), 29–47.
8. KLEENE, S. C. Representation of events in nerve nets and finite automata. *Automata Studies*, Ann. Math. Studies, No. 34, Princeton Univ. Press, 1956, 3–41.
9. POST, E. L. A variant of a recursively unsolvable problem. *Bull. Am. Math. Soc. 52*, (1946), 264–268.
10. RABIN AND SCOTT. Finite automata and their decision problems. *IBM J. Res. Develop. 3* (1959), 114–125.
11. SCHEINBERG, S. Note of the Boolean properties of context free languages. *Inform. Contr. 3* (1960), 372–375.
12. SHAMIR, E. On sequential languages. Tech. Report No. 7, Appl. Logic Branch, The Hebrew Univ., Jerusalem, Nov. 1961.