SEQUENCES IN CONTEXT FREE LANGUAGES[1]

BY

SEYMOUR GINSBURG, THOMAS N. HIBBARD, AND JOSEPH S. ULLIAN

Introduction

In [8] it was shown by a complicated argument that for two (context free) languages L_1 and L_2, it is recursively unsolvable whether there exists a complete sequential machine mapping L_1 into L_2. Now an alternative (and quite simple) proof of this fact would follow from verification of the following conjecture: It is recursively unsolvable whether a language contains an ultimately periodic sequence. (For the language $\{a^n/n \geq 1\}$ can be mapped into an arbitrary language L by a complete sequential machine if and only if L contains an ultimately periodic sequence.) This conjecture and its analogue for sequences in general are herein verified. They provide the motivation for the study of sequences in languages.

The paper is divided into three sections. Section 1 reviews the terminology of languages. In Section 2 it is shown that whether a language contains a given sequence is in general unsolvable, but that whether a language contains a given ultimately periodic sequence is solvable. The unsolvability of whether a language contains a sequence and whether a language contains an ultimately periodic sequence are also demonstrated here. Section 3 is concerned with sequences D having the property that there is a language containing D and no other sequence. (Such a sequence is called *distinguished.*) It is first shown that every distinguished sequence is recursive. Then a method of generating recursive sequences which are not distinguished is exhibited. Finally, it is shown that there are languages which contain sequences but no recursive sequence.

1. Preliminaries

Let Σ be a finite nonempty set and let $\theta(\Sigma)$ be the free semigroup with identity ε generated by Σ. (Thus $\theta(\Sigma)$ is the set of all words over Σ, and ε is the empty word.) We shall be considering certain subsets of $\theta(\Sigma)$ which are called "context free languages", or "languages" for short. These languages arose in the study of natural languages [2] and have been shown to be identical with the components in the "ALGOL-like" artificial languages which occur in data processing [6].

A *grammar* G is a 4-tuple (V, Σ, P, σ), where V is a finite set, Σ is a subset of V, σ is an element of $V - \Sigma$, and P is a finite set of ordered pairs of the form (ξ, w) with ξ in $V - \Sigma$ and w in $\theta(V)$. P is called the set of *productions*

Received January 17, 1964.

[1] This research was supported in part by Air Force Cambridge Research Laboratories.

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 exist a_i, b_i, ξ_i, w_i such that $x_i = a_i \xi_i b_i$, $x_{i+1} = a_i w_i b_i$, and $\xi_i \rightarrow w_i$. The *language generated by* G, denoted by $L(G)$, is the set of words

$$\{w/\sigma \Rightarrow w, w \text{ in } \theta(\Sigma)\}.$$

A *context free language* (over Σ) is a language $L(G)$ generated by some grammar $G = (V, \Sigma, P, \sigma)$. Unless otherwise stated, by a language we shall always mean a context free language.

If A and B are subsets of $\theta(\Sigma)$, then the set of words $\{ab/a \text{ in } A, b \text{ in } B\}$ is called the *product* of A and B and is written AB. If A (or B) consists of just one word, say $A = \{a\}(B = \{b\})$, then aB (Ab) is written instead of AB.

If A and B are languages, then so are AB, $A \cup B$, and[2] A^* [1].

The family of *regular* sets is characterized as the smallest family of subsets of $\theta(\Sigma)$ containing the finite sets and closed under the operations of union, product, and * [10]. Each regular set is a language [3].

Let $x_1, \cdots, x_n, \cdots$ (written $x_1 \cdots x_n \cdots$) be an infinite sequence of elements of Σ. A set H of words is said to *contain the sequence* $x_1 \cdots x_n \cdots$ if H contains the word $x_1 \cdots x_i$ for each i. H is said to *contain a sequence* if H contains some sequence $x_1 \cdots x_n \cdots$. Clearly, containment of a sequence corresponds to containment of a set of words closed under initial segmentation in which there is exactly one word of each positive length.

We are interested in establishing (a) the unsolvability of whether an arbitrary language contains a sequence, and (b) the unsolvability of whether an arbitrary language contains an ultimately periodic sequence.[3] We shall demonstrate (a) and (b) as well as a number of related results.

2. Solvability questions

We first consider the solvability of whether an arbitrary language contains a specific sequence. We exhibit one set of sequences for which it is unsolvable and another for which it is solvable. Then we show that it is unsolvable whether an arbitrary language contains a sequence and whether an arbitrary language contains an u.p. sequence.

LEMMA 2.1. *There exists a sequence D such that it is recursively unsolvable whether an arbitrary language over a three letter alphabet contains D.*

Proof. Let F be the set of all sequences of the form $cw_1cw_2cw_3 \cdots$, where

---

[2] If A is a set of words, then $A^* = \bigcup_0^\infty A^n$, where $A^0 = \{\varepsilon\}$ and $A^{i+1} = A^iA$ for $i \geq 0$.

[3] A sequence $x_1 \cdots x_n \cdots$ is said to be *ultimately periodic* (abbreviated u.p.) if there exist positive integers n_0 and p such that $x_{n+p} = x_n$ for $n \geq n_0$.

$\bigcup_{i=1}^{\infty} w_i = \theta(a, b)$. Given a language $M \subseteq \theta(a, b)$, consider

$$L(M) = \text{Init }[(cM)^*].^4$$

If $M = \theta(a, b)$, then $L(M)$ contains D for every D in F. If $M \neq \theta(a, b)$, then $L(M)$ contains D for no D in F. Since it is unsolvable whether $M = \theta(a, b)$ for an arbitrary language $M \subseteq \theta(a, b)$ [1], it is unsolvable whether $L(M)$,thus an arbitrary language, contains D for some specific (or even one) D in F.

THEOREM 2.1. *There exists a sequence D such that it is recursively unsolvable whether an arbitrary language over a two letter alphabet contains D.*

Proof. Let a, b, c, d, and e be distinct letters. Let τ be the function on $\theta(a, b, c)$ defined by $\tau(\varepsilon) = \varepsilon$, $\tau(a) = de$, $\tau(b) = d^2e$, $\tau(c) = d^3e$, and

$$\tau(x_1 \cdots x_k) = \tau(x_1) \cdots \tau(x_k),$$

each x_i in $\{a, b, c\}$. It is known [1] that τ preserves languages. For each letter x and each set A of words let

$$\text{Init}_x(A) = \{wx/wxy \text{ in } A \text{ for some word } y\}.$$

It is readily seen that $\text{Init}_x(A)$ is a language if A is. Let $P \subseteq \theta(a, b, c)$ be a language. Then $L(P) = \tau(P) \cup \text{Init}_d(\tau(P))$ is a language. For each sequence $D = x_1 \cdots x_n \cdots$, x_i in $\{a, b, c\}$, let $\tau(D) = \tau(x_1) \cdots \tau(x_n) \cdots$. Clearly P contains a sequence D if and only if $L(P)$ contains $\text{Init}(\tau(D))$. By Lemma 2.1, the former is unsolvable. Thus the latter is unsolvable and the theorem follows.

We next show that it is solvable whether a language contains a given u.p. sequence.

LEMMA 2.2. *Given words w_1, w_2, w_3, it is recursively solvable whether an arbitrary language L contains $w_1 w_2^* w_3$.*

Proof. If $w_2 = \varepsilon$, then it is recursively solvable whether $w_1 w_3$ is in L. Suppose that $w_2 \neq \varepsilon$. Since $w_1 w_2^* w_3$ is regular and L is a language, $A = w_1 w_2^* w_3 \cap L$ is a language and effectively calculable from L [1]. Since $w_1 w_2^* w_3 \subseteq L$ if and only if $A = w_1 w_2^* w_3$, it suffices to show that whether A and $w_1 w_2^* w_3$ are equal is solvable. Let τ_1 (τ_2) be the operation which maps a word x into x_1 if $x = w_1 x_1$ ($x = x_1 w_3$) and into φ otherwise. Then A and $w_1 w_2^* w_3$ are equal if and only if $\tau_2 \tau_1(A) = w_2^*$. Now $\tau_2 \tau_1(A)$ is a language and effectively calculable from A [7]. Since $w_2 \neq \varepsilon$, $w_2 = y_1 \cdots y_r$, y_i in Σ.

[4] For a sequence $x_1 \cdots x_n \cdots$ of symbols in Σ,

$$\text{Init}(x_1 \cdots x_n \cdots) = \{x_1 \cdots x_n / n \geq 1\}.$$

Thus a sequence D is contained in a set of words H if and only if $\text{Init}(D) \subseteq H$. For a word w, $\text{Init}(w) = \{u/u \neq \varepsilon,\ uv = w \text{ for some } v\}$. For a set H of words, $\text{Init}(H) = \bigcup_{w \text{ in } H} \text{Init}(w)$. It is known [7] that $\text{Init}(L)$ is a language if L is.

Consider the generalized sequential machine[5]

$$S = (K, \Sigma, \{a\}, \delta, \lambda, p_1),$$

where $K = \{p_1, \cdots, p_r\}$, $\lambda(p_i, y) = \varepsilon$ for $i \neq r$, $\lambda(p_r, y) = a$, $\delta(p_i, y) = p_{i+1}$ for $i < r$, and $\delta(p_r, y) = p_1$, y in Σ. Then

$$S[\tau_2\tau_1(A)] = \{a^k/w_2^k \quad \text{in} \quad \tau_2\tau_1(A)\}.^6$$

Since $\tau_2\tau_1(A)$ is a language, $S[\tau_2\tau_1(A)]$ is a language and effectively calculable from $\tau_2\tau_1(A)$ and S. From [6], a language on one letter is a regular set and is effectively calculable as a regular set. But $A = w_1 w_2^* w_3$ if and only if $S[\tau_2\tau_1(A)] = a^*$. Now it is solvable whether two regular sets are equal [10]. Thus it is solvable whether $S[\tau_2\tau_1(A)] = a^*$. Hence the result.

THEOREM 2.2 *Given an u.p. sequence D, it is solvable whether an arbitrary language contains D.*

Proof. Let D be an u.p. sequence. Then $D = w(a_1 \cdots a_p)(a_1 \cdots a_p) \cdots$, each a_i in Σ, for some word w and some $p \geq 1$. For any language L, L contains D if and only if L contains each of the following $p + 1$ sets:

$$\mathrm{Init}(w),\ w(a_1 \cdots a_p)^*,\ wa_1(a_2 \cdots a_p a_1)^*,\ wa_1 a_2(a_3 \cdots a_p a_1 a_2)^*,$$
$$\cdots,\ wa_1 \cdots a_{p-1}(a_p a_1 \cdots a_{p-1})^*.$$

It is solvable whether an arbitrary language contains $\mathrm{Init}(w)$ since $\mathrm{Init}(w)$ is finite and it is solvable whether a language contains a given word. Each of the other inclusions is solvable by Lemma 2.2. Thus whether L contains D is solvable.

We now turn to the problem of determining whether an arbitrary language contains a sequence (u.p. sequence).

Notation. Let $\varepsilon^+ = \varepsilon$ and $(x_1 \cdots x_k)^+ = x_k \cdots x_1$, each x_i in Σ.

LEMMA 2.3. *Let Σ be a (possibly infinite) alphabet. If it is decidable whether an arbitrary language whose alphabet is included in Σ contains a sequence (u.p.*

[5] A *generalized sequential machine* S is a 6-tuple $(K, \Sigma, \Delta, \delta, \lambda, p_1)$ where (i) K is a finite nonempty set (of "states"); (ii) Σ is a finite nonempty set (of "inputs"); (iii) Δ is a finite nonempty set (of "outputs"); (iv) δ is a mapping of $K \times \Sigma$ into K (the "next state" function); (v) λ is a mapping of $K \times \Sigma$ into $\theta(\Delta)$ (the "output" function); and (vi) p_1 is an element of K (the "start" state). A *complete sequential machine* is a generalized sequential machine in which λ maps $K \times \Sigma$ into Δ.

[6] Extend δ and λ to $K \times \theta(\Sigma)$ as follows. Let $\delta(q, \varepsilon) = q$ and $\lambda(q, \varepsilon) = \varepsilon$. For each word $w_1 \cdots x_{k+1}$, each x_i in Σ, let

$$\delta(q, x_1 \cdots x_{k+1}) = \delta[\delta(q, x_1 \cdots x_k), x_{k+1}]$$

and

$$\lambda(q, x_1 \cdots x_{k+1}) = \lambda(q, x_1 \cdots x_k)\lambda[\delta(q, x_1 \cdots x_k), x_{k+1}].$$

For each word w, let $S(w) = \lambda(p_1, w)$. For each set L, let $S(L) = \{S(w)/w \text{ in } L\}$. It is known that $S(L)$ is a language if L is, and is effectively calculable from L [7].

sequence), then it is solvable whether the intersection of a pair of languages whose alphabets are included in Σ contains a sequence (u.p. sequence).

Proof. Let X and Y be given languages with alphabets included in Σ, and let Σ_1 be the union of their alphabets. Let τ_1 and τ_2 be the mappings of $\theta(\Sigma_1)$ into $\theta(\Sigma_1)$ defined by $\tau_1(\varepsilon) = \tau_2(\varepsilon) = \varepsilon$, $\tau_1(x_1) = x_1$, $\tau_2(x_1) = x_1^2$, $\tau_1(x_1 \cdots x_n) = x_1^2 \cdots x_{n-1}^2 x_n$, and $\tau_2(x_1 \cdots x_n) = x_1^2 \cdots x_n^2$, $n > 1$ and each x_i in Σ_1. By [1], τ_2 preserves languages. The function τ_1 also preserves languages. For let τ_3 be the function defined by $\tau_3(\varepsilon) = \varepsilon$, $\tau_3(x_1) = x_1$, and $\tau_3(x_1 \cdots x_n) = x_1 x_2^2 \cdots x_n^2$, $n > 1$ and each x_i in Σ_1. Let S be the generalized sequential machine $(\{p_1, p_2\}, \Sigma_1, \Sigma_1, \delta, \lambda, p_1)$ with $\delta(p_1, x) = \delta(p_2, x) = p_2$, $\lambda(p_1, x) = x$, and $\lambda(p_2, x) = x^2$, x in Σ_1. Since $\tau_3(w) = S(w)$ for w in $\theta(\Sigma_1)$ and S preserves languages, τ_3 preserves languages. Now $\tau_1(M) = \tau_3(M^+)^+$, and the operation $^+$ preserves languages [1]. Thus τ_1 preserves languages.

Since τ_1 and τ_2 preserve languages, $\tau_1(X) \cup \tau_2(Y)$ is a language with alphabet a subset of Σ. Clearly $X \cap Y$ contains a sequence (u.p. sequence) if and only if $\tau_1(X) \cup \tau_2(Y)$ contains a sequence (u.p. sequence). If it is decidable whether an arbitrary language whose alphabet is included in Σ, hence $\tau_1(X) \cup \tau_2(Y)$, contains a sequence (u.p. sequence), then it is solvable whether $X \cap Y$ contains a sequence (u.p. sequence).

In the next three lemmas and Theorem 3.3 we shall use the terminology and notation of Turing machines as formulated in [4, pp. 5–7]. Thus we shall speak of the alphabet of a Turing machine Z, instantaneous descriptions u, v, $u \rightarrow v(Z)$, etc.

Notation. Let Z be a Turing machine. Write $v = Z(u)$ if $u \rightarrow v(Z)$. Let $Z^0(u) = u$ and $Z^{i+1}(u) = Z(Z^i(u))$, provided $Z(Z^i(u))$ exists.

LEMMA 2.4. *Let Z be a Turing machine and c a letter not occurring in any instantaneous description of Z. Then the set $\{u^+cv/v = Z(u)\}$ is a language.*

Proof. Let $G = (V, \Sigma, P, \sigma)$ be the grammar defined as follows. Σ is the alphabet of Z together with the internal configurations of Z together with c. $V - \Sigma = \{\sigma, \xi_1, \xi_2, \xi_3\}$. P consists of those productions having the following form:

(1) $\sigma \rightarrow \xi_1$, $\sigma \rightarrow \xi_3$.
(2) $\xi_1 \rightarrow S_i \xi_1 S_i$ for each symbol S_i in the alphabet of Z.
(3) $\xi_1 \rightarrow S_j q_i \xi_2 q_m S_k$ whenever (q_i, S_j, S_k, q_m) is in Z.
(4) $\xi_1 \rightarrow S_k S_j q_i \xi_2 S_j q_m S_k$ for each S_k in the alphabet of Z, whenever (q_i, S_j, R, q_m) is in Z.
(5) $\xi_3 \rightarrow S_j q_i \xi_2 S_j q_m S_0$ whenever (q_i, S_j, R, q_m) is in Z.
(6) $\xi_1 \rightarrow S_j q_i S_k \xi_2 q_m S_k S_j$ for each S_k in the alphabet of Z, whenever (q_i, S_j, L, q_m) is in Z.
(7) $\xi_1 \rightarrow S_j q_i c q_m S_0 S_j$ whenever (q_i, S_j, L, q_m) is in Z.

(8) $\xi_2 \to S_i \, \xi_2 \, S_i$ for each S_i in the alphabet of Z.
(9) $\xi_2 \to c$.

It is a straightforward matter to verify that $L(G) = \{u^+cv/v = Z(u)\}$.

LEMMA 2.5. *Let Z be a Turing machine and c a letter not occurring in any instantaneous description of Z. Then for each instantaneous description w of Z, there exist languages A and B such that*

$$A \cap B = \text{Init}(v_0 v_1 v_2 \cdots),$$

where $v_i = Z^i(w)c(Z^i(w))^+c$ if $Z^i(w)$ exists and $v_i = \varepsilon$ if $Z^i(w)$ does not exist.

Proof. Let $A_1 = \{u^+cv/v = Z(u)\}$. By Lemma 2.4, A_1 is a language. Let I be the set of instantaneous descriptions of Z and $A_2 = \{ucu^+/u \text{ in } I\}$. Now an instantaneous description is an expression that contains exactly one internal configuration q_i, neither of the symbols R or L, and is such that q_i is not the right-most symbol. Thus

$$I = \bigcup_i \{S_0, \cdots, S_s\}^* q_i \{S_0, \cdots, S_s\}\{S_0, \cdots, S_s\}^*,$$

where $S_0, \cdots, S_s$ is the alphabet of Z. Hence I is regular. It is known that $\{ucu^+/u \text{ in } I\}$ is a language if I is regular [1]. Hence A_2 is a language. Let

$$A = \text{Init}(wc(A_1 c)^*) \quad \text{and} \quad B = \text{Init}((A_2 c)^*).$$

Since Init preserves languages, A and B are languages. To complete the proof we shall show that $M = A \cap B$, where $M = \text{Init}(v_0 v_1 \cdots)$.

Clearly $M \subseteq A \cap B$. Consider the reverse inclusion. First note that

(1) if $u_0 c u_1 c \cdots u_{2m} c u_{2m+1} c$ is in B $(m \geq 0)$, where no u_j contains an occurrence of c, then $u_{2i+1} = \overset{+}{u_{2i}}$ for each $i \geq 0$.

Next note that

(2) if $u_0 c u_1 c \cdots u_{2m} c$ is in A $(m \geq 0)$, where no u_j contains an occurrence of c, then $u_0 = w$ and $u_{2i} = Z(\overset{+}{u_{2i-1}})$ for each $i \geq 1$.

Now let u be an element of $A \cap B$. If u contains no occurrence of c, then obviously u is in $\text{Init}(w)$ and thus in M. Suppose that u contains an even number $2m > 0$ of occurrences of c. Let $u = u_0 c u_1 c \cdots u_{2m-1} c u_{2m}$. Since u is in B, so is $u_0 c u_1 c \cdots u_{2m-1} c u_{2m}$. Since u is in B, so is $u_0 c u_1 c \cdots u_{2m-1} c$. By (1), $u_{2i+1} = \overset{+}{u_{2i}}$ for each i $(0 \leq i < m)$. Since u is in A, there exists a word y containing no occurrence of c such that uyc is in A. By (2), $u_0 = w$, $u_{2i} = Z(\overset{+}{u_{2i-1}})$ for $1 \leq i < m$, and $u_{2m} y = Z(\overset{+}{u_{2m-1}})$. Thus

$$uyc = wcw^+cZ(w)c(Z(w))^+c \cdots Z^m(w)c(Z^m(w))^+c.$$

Hence uyc is in M, so that u is in M. A parallel argument arises when u contains an odd number of occurrences of c. Thus $A \cap B \subseteq M$ and the lemma is established.

LEMMA 2.6. *Let E be the family of languages with alphabet included in $\Sigma = \{c, S_0, S_1, \cdots, q_1, q_2, \cdots\}$. It is recursively unsolvable whether an arbitrary language in E*

(1) *contains a sequence;*

(2) *contains an u.p. sequence.*

Proof. Let Z be a Turing machine, c a letter not occurring in any instantaneous description of Z, and w an instantaneous description of Z. Let A, B be the languages of Lemma 2.5, and $D = v_0 v_1 \cdots$. Then $A \cap B$ contains a sequence if and only if D is infinite, that is, there is no m such that $v_j = \varepsilon$ for all $j \geq m$. D is infinite if and only if $Z^i(w) \neq \varepsilon$ for all i, that is, if and only if Z, starting at w, does not halt.

Suppose that (1) is solvable. Then by Lemma 2.3, it is decidable whether $A \cap B$ contains a sequence, whence decidable whether Z halts starting at w. Since the halting problem is not decidable, (1) is unsolvable.

Now $A \cap B$ contains an u.p. sequence if and only if D is u.p. D is u.p. if and only if there exist i and j, $i < j$, such that $Z^i(w) \neq \varepsilon$ and $Z^i(w) = Z^j(w)$, that is, if and only if Z, starting at w, has a repeating instantaneous description. Suppose that (2) is solvable. Then it is solvable whether Z, starting at w, has a repeating instantaneous description. Since the repeating problem is unsolvable, (2) is unsolvable.

THEOREM 2.3. *It is recursively unsolvable whether an arbitrary language over two letters*

(1) *contains a sequence;*

(2) *contains an u.p. sequence.*

Proof. Let d and e be distinct letters. Let E be the family of languages with alphabet included in $\Sigma = \{c, S_0, S_1, \cdots, q_1, q_2, \cdots\}$. For each n let $\Sigma_n = \{c, S_0, \cdots, S_n, q_1, \cdots, q_n\}$ and τ_n be the function on $\theta(\Sigma_n)$ defined by $\tau_n(\varepsilon) = \varepsilon$, $\tau_n(c) = de$, $\tau_n(S_i) = d^{2i+2}e$ for $0 \leq i \leq n$, $\tau_n(q_i) = d^{2i+1}e$ for $1 \leq i \leq n$, and $\tau_n(x_1 \cdots x_k) = \tau_n(x_1) \cdots \tau_n(x_k)$, each x_i in Σ_n. The function τ_n preserves languages. For each language $P \subseteq \theta(\Sigma_n)$ let

$$L(P) = \tau_n(P) \cup \mathrm{Init}_d(\tau_n(P)),$$

where

$$\mathrm{Init}_d(A) = \{wd/wdy \text{ in } A \text{ for some word } y\}.$$

As in Theorem 2.1, $L(P)$ is a language and P contains a sequence D if and only if $L(P)$ contains $\mathrm{Init}(\tau_n(D))$. Furthermore, $L(P)$ contains a sequence D' if and only if $D' = \tau_n(x_1) \cdots \tau_n(x_j) \cdots$ for some sequence $x_1 \cdots x_j \cdots$ in P.

Suppose it is solvable whether an arbitrary language M over $\{d, e\}$ contains a sequence. Then it is solvable whether an arbitrary language in E contains a sequence, contradicting Lemma 2.6. Hence it is unsolvable whether an arbitrary language over $\{d, e\}$ contains a sequence.

The recursive unsolvability of whether a language over $\{d, e\}$ contains an u.p.

sequence follows fom Lemma 2.6 and the fact that a sequence $x_1 \cdots x_j \cdots$ is u.p. if and only if $\tau_n(x_1) \cdots \tau_n(x_j) \cdots$ is u.p.

Remarks. (1) Properties of languages are usually shown to be undecidable by reduction to the Post Correspondence Theorem, that is, the unsolvability of determining for arbitrary n-tuples $(w_1, \cdots, w_n)$ and $(y_1, \cdots, y_n)$ of words over $\{a, b\}$, whether there exist $i_1, \cdots, i_k$ such that $w_{i_1} \cdots w_{i_k} = y_{i_1} \cdots y_{i_k}$ [9]. We now outline an alternative proof of Lemma 2.6 which depends on the Post Correspondence Theorem. Here $\Sigma = \{a, b, c\}$ for the sequence case and $\Sigma = \{a, b, c, d, a_1, a_2, \cdots, a_n, \cdots\}$ for the u.p. sequence case.

First consider the sequence problem. Let T be the "successor" function defined on $\theta(a, b)$ by $T(b^n) = a^{n+1}$ and $T(wab^n) = wba^n$, $n \geq 0$ and w in $\theta(a, b)$. Thus T "enumerates" $\theta(a, b)$ from ε as follows: ε, a, b, aa, ab, ba, bb, aaa, aab, $\cdots$. Let

$$A_1 = \{w^+cT(w)/w \text{ in } \theta(a, b)\}, \qquad A_2 = \{wcw^+/w \text{ in } \theta(a, b)\},$$

$$A = \text{Init}[cc(A_1 c)^*], \quad \text{and} \quad B = \text{Init}[c(A_2 c)^*].$$

For $i \geq 0$ let $v_i = T^i(\varepsilon)c(T^i(\varepsilon))^+$, where $T^0(\varepsilon) = \varepsilon$ and $T^{j+1}(\varepsilon) = T(T^j(\varepsilon))$. Then $A \cap B = \text{Init}(D)$, where $D = cv_0 cv_1 \cdots cv_n \cdots$. By the proof of Lemma 2.1, it is unsolvable whether an arbitrary language over $\{a, b, c\}$ contains D. (The proof of Lemma 2.1 reduces to the unsolvability of determining, for a language $M \subseteq \theta(a, b)$, whether $M = \theta(a, b)$. This in turn reduces to the Post Correspondence Theorem [1].) For any language X, $A \cap B \cap X$ contains a sequence if and only if X contains D. By an argument similar to that in Lemma 2.3, this implies the unsolvability of the sequence problem over a three letter alphabet.

Now consider the u.p. sequence problem. For each j let T_j be the obvious "successor" function over $\theta(a_1, \cdots, a_j)$. Let $(w_1, \cdots, w_n)$ and $(y_1, \cdots, y_n)$ be given n-tuples $(n \geq 1)$ of non-ε words in $\theta(a, b)$. Let

$$Y = \{a_{i_1} \cdots a_{i_p} ca_{j_q} \cdots a_{j_1}/w_{i_1} \cdots w_{i_p} = y_{j_1} \cdots y_{j_q}\}$$

and

$$Z = \{a_{i_1} \cdots a_{i_p} ca_{j_q} \cdots a_{j_1}/w_{i_1} \cdots w_{i_p} \neq y_{j_1} \cdots y_{j_q}\}.$$

Y and Z are languages over $\{a_1, \cdots, a_n, c\}$. Let

$$A_1 = \{w^+cT_n(w)/w \text{ in } \theta(a_1, \cdots, a_n)\},$$

$$A_2 = \{wcw^+/w \text{ in } \theta(a_1, \cdots, a_n)\},$$

$$A = \text{Init}\,[a_1 c(A_1 c)^* \theta(a_1, \cdots, a_n)\, da_1^*],$$

$$B = \text{Init}\,[(A_2 c)^* A_2\, da_1^*],$$

$$F = \text{Init}\,[(Zc)^* Y\, da_1^*].$$

$A \cap B \cap F$ contains an u.p. sequence if and only if Y contains a word ucu^+, u in $\theta(a_1, \cdots, a_n) - \varepsilon$. Y contains such a word if and only if there exist

$i_1, \cdots, i_p$ such that $w_{i_1} \cdots w_{i_p} = y_{i_1} \cdots y_{i_p}$, which is recursively unsolvable.

(2) The following problems may be shown to be recursively unsolvable either by extension of the methods of Lemma 2.6, 2.3, and Theorem 2.3 or by reduction to the Post Correspondence Theorem: Does an arbitrary language over a two letter alphabet

(a) contain a purely periodic sequence?[7]

(b) contain, for a given word $x \neq \varepsilon$, an u.p. sequence with period x, that is, a sequence of the form $x_1 xx \cdots$?

Theorem 2.3 permits us to obtain a result about "counting chains."

THEOREM 2.4. *Call $C(P)$ a counting chain, where P is a subset of the positive integers, if $C(P)$ is the set of those words*

$$b^{i_1}a^{i_1}b^{i_2}a^{i_1+i_2} \cdots b^{i_k}a^{\Sigma_{j=1}^k i_j},$$

$k \geq 1$, *such that for* $1 \leq j \leq k$, $i_j = 2$ *if j is in P and* $i_j = 1$ *if j is not in P. Then the question of whether an arbitrary language over $\{a, b\}$ contains a counting chain is recursively unsolvable.*

Proof. Let σ be the operation which takes each occurrence of b into $\{b, b^2\}$ and leaves a unchanged. That is, $\sigma(\varepsilon) = \varepsilon$ and $\sigma(x_1 \cdots x_r) = \sigma(x_1) \cdots \sigma(x_r)$, where $\sigma(a) = a$ and $\sigma(b) = \{b, b^2\}$. Let τ be the operation which takes each occurrence of a into $A_1 = \{ba^{2n}/n \geq 1\}$ and each occurrence of b into $A_2 = \{ba^{2n+1}/n \geq 0\}$. Let σ' and τ' be the operations defined by $\sigma'(H) = \bigcup_{h \text{ in } H} \sigma(h)$ and $\tau'(H) = \bigcup_{h \text{ in } H} \tau(h)$ respectively. Since A_1 and A_2 are languages, σ' and τ' preserve languages, i.e., if M is a language so are $\sigma'(M)$ and $\tau'(M)$ [1]. We shall show that for a language $M \subseteq \theta(a, b), \sigma'\tau'(M)$ contains a counting chain if and only if M contains a sequence. Since it is unsolvable whether M contains a sequence, the theorem will follow.

To this end let M be a subset of $\theta(a, b)$. Suppose that M contains a sequence $D = x_1 \cdots x_n \cdots$, each x_i in $\{a, b\}$. For each n let $d_n = x_1 \cdots x_n$. Let $A_3 = \{1\}$ if $x_1 = a$ and $A_3 = \varphi$ if $x_1 = b$. Let

$$P = A_3 \cup \{n/n \geq 2, x_{n-1} = x_n\}.$$

For each n, $ba^{i_1}ba^{i_2}b \cdots ba^{i_n}$ is in $\tau(d_n)$ if and only if $i_j > 0$ for $1 \leq j \leq n$ and

$$\{j/j \leq n, i_j \text{ even}\} = \{j/j \leq n, x_j = a\}.$$

For each n let u_n be the element in $\tau(d_n)$ with $i_1 = 1$ or 2 and $i_{j+1} = i_j + 1$ or $i_{j+1} = i_j + 2$, $1 \leq j < n$. Then

$$\sigma(u_n) = \{b^{k_1}a^{i_1}b^{k_2} \cdots b^{k_n}a^{i_n}/k_j = 1, 2; 1 \leq j \leq n\}.$$

In particular, $\sigma(u_n)$ contains an element v_n in which $k_j = 2$, $1 \leq j \leq n$, if and only if j is in P. Thus $k_1 = 2$ if and only if $x_1 = a$. Hence $k_1 = i_1$. Further-

[7] An infinite sequence $x_1 \cdots x_n \cdots$ is said to be *purely periodic* if there exists an integer $m \geq 1$ such that $x_{i+m} = x_i$ for all $i \geq 1$.

more, for each j, $k_{j+1} = 2$ if and only if $x_j = x_{j+1}$, whence $k_{j+1} = i_{j+1} - i_j$. Thus $\{v_i/i \geq 1\}$ is the counting chain $C(P)$. Since

$$\{v_i/i \geq 1\} \subseteq \sigma'(\{u_n/n \geq 1\}),$$

$$\{u_n/n \geq 1\} \subseteq \tau'(\mathrm{Init}(D)), \quad \text{and} \quad \mathrm{Init}(D) \subseteq M,$$

it follows that $C(P) \subseteq \sigma'\tau'(M)$.

Now suppose that the counting chain $C(P) \subseteq \sigma'\tau'(M)$ for some set P of positive integers. For each n let

$$v_n = b^{i_1}a^{i_1} \cdots b^{i_n}a^{\sum_{j=1}^{n} i_j}$$

be in $C(P)$. Then there is a unique word d_n in M such that v_n is in $\sigma'\tau(d_n)$. In fact, d_n is the word $x_1 \cdots x_n$ of length n in $\theta(a, b)$ for which (i) $x_1 = a$ if and only if $i_1 = 2$, and (ii) for $1 \leq j < n$, $x_j = x_{j+1}$ if and only if $i_{j+1} = 2$. It readily follows that $\{d_i/i \geq 1\} = \mathrm{Init}(D)$ for some sequence D, and D is a sequence in M.

3. Distinguished sequences

Consider the question of whether or not a language containing a sequence must contain an u.p. sequence. By application of a systematic procedure, we can effectively enumerate those languages which contain no sequences. Since we can test a given language to see if it contains a specified u.p. sequence (Theorem 2.2), and since we can effectively enumerate the u.p. sequences, we also have a systematic procedure for effectively enumerating those languages which contain u.p. sequences. Therefore, if each language containing a sequence contained an u.p. sequence, we would have a decision procedure for determining whether or not an arbitrary language contained a sequence. By Theorem 2.3, there is no such decision procedure. Hence there exists a language which contains a sequence but no u.p. sequence. This is established constructively by the following example.

Example. Given a word w and an element b in Σ, let $\# b(w)$ be the number of occurrences of b in w.

Let $G_1 = (V_1, \Sigma, P_1, \xi)$, where $\Sigma = \{a, b\}$, $V_1 = \xi \cup \Sigma$, and $P_1 = \{\xi \to b, \xi \to a\xi, \xi \to b\xi a\}$. Let $L_1 = L(G_1)$. Clearly

$$L_1 = \{u/u = wba^{\# b(w)}, w \text{ in } \theta(a, b)\}.$$

If $u = wba^n$ is in L_1, then wba^nba^{n+1} is in L_1 and is the proper extension of u in L_1 of smallest length. Hence L_1 contains the set $\{b, bba, bbaba^2, bbaba^2ba^3, \cdots\}$.

Let $G_2 = (V_2, \Sigma, P_2, \xi)$, where $V_2 - \Sigma = \{\xi, \nu, \gamma\}$ and $P_2 = \{\xi \to \nu\gamma, \gamma \to ba, \gamma \to a\gamma, \gamma \to b\gamma, \gamma \to b\gamma a, \nu \to b, \nu \to a\nu, \nu \to b\nu, \nu \to \nu a\}$. Let $L_2 = L(G_2)$. Then

$$L_2 = \{u/u = wba^n, 1 \leq n \leq \# b(w), w \text{ in } \theta(a, b)b\theta(a, b)\}.$$

Let $L_3 = \{b\} \cup L_1 b \cup L_2$. Note that each word in $\{b\} \cup L_1 b$ ends in b, and each word in L_2 ends in a. Let D be the sequence $bbaba^2ba^3b \cdots$. Obviously D is not u.p. We shall show that L_3 contains D but no other sequence.

Each word in Init(D) ending in b is in $\{b\} \cup L_1 b$. Since each word in Init(D) ending in a is in L_2, L_3 contains D. Now let E be any sequence contained in L_3. Neither a nor ba is in L_2, thus neither is in L_3. Therefore E begins with bb. Now suppose that E begins with $buba^n$, $n \geq 0$, for some word u in $\theta(a, b)$. Two cases arise.

(α) $n = \# b(bu)$. Then $buba^{n+1}$ is not in L_2, and thus not in L_3. Hence E must begin with $buba^n b$.

(β) $n < \# b(bu)$. Since $n \neq \# b(bu)$, $buba^n$ is not in L_1. Thus E cannot begin with $buba^n b$; that is, E begins with $buba^{n+1}$.

By induction it therefore follows that $E = D$. Hence D is the only sequence contained in L_3.

We now consider sequences D with the property that there is a language containing D and no other sequence.

Definition. A sequence D with the property that there is a language containing D and no other sequence is called a *distinguished* sequence.

Since Init(D) is a language for a sequence D if and only if D is u.p., each u.p. sequence is distinguished. The sequence D in the above example shows that the converse is not true, i.e., there are distinguished sequences which are not u.p.

Given a distinguished sequence D we may obtain other distinguished sequences as follows. Let S be any complete sequential machine with the property that at each state, λ maps Σ one to one into Δ. Let L be a language containing a distinguished sequence D. Then $S(L)$ is a language containing the sequence $S(D)$. That $S(L)$ contains no sequence but $S(D)$ follows from the fact that λ maps Σ one to one into Δ. Furthermore, if D is not u.p., neither is $S(D)$. We omit the straightforward details.

The question naturally arises: Are there any sequences which are not distinguished? A simple cardinality argument shows that there are. For there are $2^{\aleph_0}$ sequences when Σ contains at least two elements, and only $\aleph_0$ languages. Thus there exists a sequence D (in fact $2^{\aleph_0}$) such that any language containing D contains at least one other sequence, i.e., a sequence D which is not distinguished. In fact

Theorem 3.1. *Every distinguished sequence is recursive.*

This follows from the well known folk theorem that if a recursive tree with finite branching has a unique infinite path, then the path is recursive.

The next theorem shows the existence of recursive sequences which are not distinguished.

Theorem 3.2. *Let a be a given element of Σ. Then each recursive, non-u.p.*

sequence D with the property that for every $n \geq 1$ there is a word ua^k, $u \neq \varepsilon$, in Init(D) *such that $k \geq 2^{n|u|}$ is not distinguished,*[8] *$|u|$ denoting the length of u.*

Proof. We first recall some terminology and facts about generation trees. Let $G = (V, \Sigma, P, S)$ be a grammar. Call the elements of $V - \Sigma$ *variables.* Let w_1 be a variable. Let $w_2, \cdots, w_r$ be words in $\theta(V)$, $w_1 \rightarrow w_2$ a production, with the following property. For $2 \leq i < r$ there exist words u_i, v_i, y_i, z_i such that $w_i = u_i y_i v_i$, $w_{i+1} = u_i z_i v_i$, and $y_i \rightarrow z_i$ is a production. A generation tree (constructed below) is a rooted, directed tree with an element of $V \cup \{\varepsilon\}$, called the node name, associated at each node.

The nodes of the tree are certain tuples of the form $(i_1, \cdots, i_k)$, where $k \leq r$ and i_j is a positive integer. The directed lines of the tree are all the ordered pairs $\langle(i_1, \cdots, i_k), (i_1, \cdots, i_k, i_{k+1})\rangle$ of nodes. Let the 1-tuple (1) be the root and w_1 the node name of (1). If $w_2 = \varepsilon$ let $(1, 1)$ be a node in the tree and ε the node name of $(1, 1)$. If $w_2 = x_1^{(2)} \cdots x_{n(2)}^{(2)}$, each $x_i^{(2)}$ in V, let $(1, i)$, $1 \leq i \leq n\ (2)$, be a node and $x_i^{(2)}$ its node name. Continuing by induction, suppose that for all $t \leq k$ every occurrence in w_t of an element of V serves as node name of some node. Now

$$(*) \qquad u_k y_k v_k = w_k \Rightarrow w_{k+1} = u_k z_k v_k .$$

Let $(i_1, \cdots, i_s)$ be the node whose node name is the occurrence of y_k indicated in $(*)$. If $z_k = \varepsilon$ let $(i_1, \cdots, i_s, 1)$ be a node and ε its node name. If $z_k = x_1^{(k)} \cdots x_{n(k)}^{(k)}$, each $x_i^{(k)}$ in V, let $(i_1, \cdots, i_s, i)$, $1 \leq i \leq n(k)$, be a node and $x_i^{(k)}$ its node name. This procedure is repeated through $k = r - 1$. The resulting entity is the generation tree.

A node $(j_1, \cdots, j_t)$ is said to be an *extension* of the node $(i_1, \cdots, i_s)$ if $s \leq t$ and $i_k = j_k$ for all $k \leq s$.

A *path* in a generation tree is a sequence of nodes $N_1, \cdots, N_k$ such that $\langle N_i, N_{i+1}\rangle$ is a directed line for each $i \leq k - 1$.

Given the nodes $N_1 = (i_1, \cdots, i_s)$ and $N_2 = (j_1, \cdots, j_t)$ write $N_1 \leq N_2$ if either N_2 is an extension of N_1 or if $i_k < j_k$ for the smallest integer k such that $i_k \neq j_k$.

The relation $\leq$ is a simple order on the set of nodes.

A node is called *maximal* if there is no node distinct from it which is an extension of it.

We shall use (implicitly and explicitly) the following known facts about a generation tree T associated with $\xi \Rightarrow w$ [1]:

(a) If N is a nonmaximal node, then the node name x of N is a variable and $\xi \Rightarrow uxv$ for some u and v in $\theta(V)$.

(b) Let $N_1, \cdots, N_k$ be the maximal nodes, with $N_i \leq N_{i+1}$ for each i.

---

[8] One such sequence $D = x_1 \cdots x_n \cdots$ is obtained by letting $f(0) = 1$, $f(n + 1) = f(n) + 2^{(n+1)f(n)} + 1$ for $n \geq 0$, $x_i = b$ if i is in the range of f, and $x_i = a$ otherwise.

Then w is the word obtained by replacing in $N_1 \cdots N_k$ each node with its node name.

(c) Let N be a nonmaximal node in a generation tree, and x its node name. Then the "subtree" of T formed by using as nodes all extensions of N is a generation tree.

(d) Let $w = u\gamma v$ and let T_1 be a generation tree of $\gamma \Rightarrow w_1$. If T_1 is placed (in the obvious way) with its root on the node whose node name is γ in $u\gamma v$, then a generation tree of $\xi \Rightarrow uw_1 v$ is obtained.

We now turn to the proof of Theorem 3.2. Let D be a sequence satisfying the hypothesis of the theorem. Let L be any language containing D. We shall show that L contains an u.p. sequence. Consider the set

$$L' = L - \{\varepsilon\} - \Sigma.$$

L' is a language and there is a grammar $G = (V, \Sigma, P, \sigma)$, $L(G) = L'$, such that every production in P is of the form $\xi \to \mu\nu$, μ and ν in V [1]. Let N denote the number of distinct variables. Let H be the set of those variables ξ such that $\xi \Rightarrow a^s\xi a^t$ for some $s + t > 0$. Let H_1 be the set of those ξ in H such that $\xi \Rightarrow \xi a^t$ for some $t > 0$. (We can effectively determine H and H_1, but we do not need this fact.) We shall see below that H is nonempty. Denote the distinct elements of H by $\xi_1, \cdots, \xi_r$. For each ξ_i in H_1 let $e(i) > 0$ be an integer such that $\xi_i \Rightarrow \xi_i a^{e(i)}$. For each ξ_i in $H - H_1$ let $e(i) > 0$, $s(i)$, $t(i)$ be integers such that

$$e(i) = s(i) + t(i) \quad \text{and} \quad \xi_i \Rightarrow a^{s(i)}\xi_i a^{t(i)}.$$

Let $e = e(1) \cdots e(r)$.

Consider any word ua^k in $\mathrm{Init}(D)$, where $u \neq \varepsilon$ and $k \geq 2^{(2N+e)|u|}$. We shall show that $\mathrm{Init}\,(ua^*) \subseteq L$, thereby proving the theorem. Since

$$\mathrm{Init}\,(ua^k) \subseteq \mathrm{Init}(D) \subseteq L,$$

it suffices to show that ua^q is in $L(G)$ for each $q > k$. Accordingly, let $q > k$ be given and let $p = |u|$. Then

$$k - 2^{2Np} \geq 2^{(2N+e)p} - 2^{2Np} = 2^{2Np}(2^{ep} - 1) \geq 2(2^{ep} - 1) \geq 2^{ep} > ep \geq e.$$

Therefore there is a positive integer g such that $2^{2Np} < q - ge \leq k$. Then ua^{q-ge} is in $\mathrm{Init}\,(ua^k)$ and $|ua^{q-ge}| \geq 2$. Thus ua^{q-ge} is in $L(G)$. Hence there is a generation tree T of g which derives ua^{q-ge} (from σ).

Since each production is of the form $\xi \to \mu\nu$, μ and ν in V, it is readily seen that any generation tree of G of a word of length $> 2^n$ contains a path with at least $n + 1$ nodes, where each node name is a variable. Now

$$|ua^{q-ge}| > q - ge > 2^{2Np} \geq 2^{N(p+1)}.$$

Thus T contains a path $Z_1, \cdots, Z_{N(p+1)+1}$, where the node name of each Z_i is a variable. Since there are only N distinct variables, one of them, say ξ, is the node name of at least $p + 2$ nodes. Denote by $Y_1, \cdots, Y_{p+2}$ the first $p + 2$ nodes in the path whose node name is ξ. For $1 \leq i \leq p + 2$, let T_i be the subtree of T whose nodes are the extensions of Y_i. Then T_i is a generation tree (from ξ) of a word v_i in $\theta(\Sigma) - \varepsilon$. For $1 \leq i \leq p+1$, since the node Y_{i+1} occurs in T_i, there are words x_i, y_i in $\theta(\Sigma)$ such that $\xi \Rightarrow x_i \xi y_i$ and $v_i = x_i v_{i+1} y_i$. Since each production is of the form $\gamma \rightarrow \mu\nu$, μ and ν in V, $x_i y_i \neq \varepsilon$. Since Y_1 is in T, there exist w_1, w_2 in $\theta(V)$ such that $\sigma \Rightarrow w_1 \xi w_2$. Thus $ua^{q-ge} = w_1 x_1 \cdots x_{p+1} v_{p+2} y_{p+1} \cdots y_1 w_2$.

Two cases arise.

(1) Suppose that one of the x_i is ε. Let j be the smallest integer such that $x_j = \varepsilon$. Then $| x_1 \cdots x_{j-1} | \geq j - 1$. Since $x_i y_i \neq \varepsilon$ for each i,

$$| x_{j+1} \cdots x_{p+1} v_{p+2} y_{p+1} \cdots y_{j+1} | \geq p + 1 - j.$$

Thus $| x_1 \cdots y_{j+1} | \geq p$, so that u is an initial subword of $w_1 x_1 \cdots y_{j+1}$. Therefore y_j is in a^*. As $x_j = \varepsilon$, $y_j \neq \varepsilon$. Thus y_j is in aa^*. Since $\xi \Rightarrow x_j \xi y_j$, ξ is in H_1, say $\xi = \xi_d$. Now e is a multiple of $e(d)$. Thus

$$\xi \Rightarrow \xi a^{e(d)ge/e(d)} = \xi a^{ge}$$

and

$$\begin{aligned}\sigma &\Rightarrow w_1 x_1 \cdots x_j \xi y_j \cdots y_1 w_2 \\ &\Rightarrow w_1 x_1 \cdots x_j \xi a^{ge} y_j \cdots y_1 w_2 \\ &\Rightarrow w_1 x_1 \cdots x_{p+1} v_{p+2} y_{p+1} \cdots y_{j+1} a^{ge} y_j \cdots y_1 w_2 \\ &= ua^{q-ge} a^{ge} = ua^q.\end{aligned}$$

(2) Suppose that none of the x_i is ε. Then $| x_1 \cdots x_p | \geq p$, so that u is an initial subword of $w_1 x_1 \cdots x_p$. Thus $x_{p+1} v_{p+2} y_{p+1} \cdots y_1 w_2$ is in aa^*. Then $\xi \Rightarrow x_{p+1} \xi y_{p+1}$, with $x_{p+1} y_{p+1}$ in aa^*. Therefore ξ is in H, say $\xi = \xi_d$. Then there exist nonnegative integers s and t so that $\xi \Rightarrow a^s \xi a^t$ and $e(d) = s + t$. Thus

$$\begin{aligned}\sigma &= w_1 x_1 \cdots x_p \xi y_p \cdots y_1 w_2 \\ &\Rightarrow w_1 x_1 \cdots x_p a^{sge/e(d)} \xi a^{tge/e(d)} y_p \cdots y_1 w_2 \\ &\Rightarrow w_1 x_1 \cdots x_p a^{sge/e(d)} x_{p+1} v_{p+2} y_{p+1} a^{tge/e(d)} y_p \cdots y_1 w_2 \\ &= w_1 x_1 \cdots x_{p+1} v_{p+2} y_{p+1} \cdots y_1 w_2 a^{(s+t)ge/e(d)} \\ &\qquad\qquad (\text{since } x_{p+1} \cdots y_1 w_2 a^{(s+t)ge/e(d)} \text{ is in } aa^*) \\ &= ua^q.\end{aligned}$$

Finally we establish the existence of a language which contains a sequence but no recursive sequence.

LEMMA 3.1. *There exists a language which contains a sequence but no recursive sequence.*

Proof. By [5], [11] there exists a recursive subset M of $\theta(a, b)$ which con-

contains a sequence but no recursive sequence. Since M is recursive we can construct a Turing machine Z with alphabet $a, b, S_0, \cdots, S_h$ ($h \geq 0$) and internal configurations $q_1, \cdots, q_m$ ($m \geq 3$) which has the following property: Suppose that Z starts from an instantaneous description $S_0^n q_1 w S_0^p$ with $n, p \geq 0$ and w in $\theta(a, b)$. Then

(1) if w is not in M, Z halts in the internal configuration q_2.

(2) if w is in M, Z halts in the instantaneous description $S_0^r q_3 w S_0^t$ for some $r \geq 0$ and $t \geq 0$.

We extend the relation $\rightarrow$ defined by the Turing machine Z. Let $\bar{a}$, $\bar{b}$, and q_{m+1} be distinct elements not in $\{a, b, S_0, \cdots, S_h, q_1, \cdots, q_m\}$. For all $r \geq 0$, $t \geq 0$, and w in $\theta(a, b)$ let

(3) $S_0^r q_3 w S_0^t \rightarrow S_0^r q_{m+1} w\bar{a} S_0^t$,

(4) $S_0^r q_3 w S_0^t \rightarrow S_0^r q_{m+1} w\bar{b} S_0^t$,

(5) $S_0^r q_{m+1} w\bar{a} S_0^t \rightarrow S_0^r q_1 wa S_0^t$,

(6) $S_0^r q_{m+1} w\bar{b} S_0^t \rightarrow S_0^r q_1 wb S_0^t$.

Let $\Sigma = \{a, b, \bar{a}, \bar{b}, S_0, \cdots, S_h, q_1, \cdots, q_{m+1}\}$. Suppose that $u_1, u_2, \cdots$ is an infinite sequence of words in $\theta(\Sigma)$ such that $u_1 = q_3$ and $u_i \rightarrow u_{i+1}$ for all i. If $x_j = a$ ($x_j = b$) let $\bar{x}_j = \bar{a}$ ($\bar{x}_j = \bar{b}$), all $j \geq 1$. Then $u_2 = q_{m+1} \bar{x}_1$, with x_1 in $\{a, b\}$ and $u_3 = q_1 x_1$. Moreover,

(7) x_1 is in M.

For suppose the contrary. By (1), u_3 (uniquely) leads to an instantaneous description containing q_2 which cannot be in the domain of the extended $\rightarrow$ relation, so that $u_1, u_2, \cdots$ terminates, a contradiction. Continuing by induction, suppose that $x_1 \cdots x_r$ is a word in M, each x_j in $\{a, b\}$, such that $u_k = S_0^n q_1 x_1 \cdots x_r S_0^p$ for some $k \geq 1$, $n \geq 0$, $p \geq 0$. By (2)–(6), there is a smallest integer $s > k$ such that $u_s = S_0^r q_3 x_1 \cdots x_r S_0^t$ for some $r, t \geq 0$. Note that neither $\bar{a}$ nor $\bar{b}$ occurs in any word $u_k, \cdots, u_s$. Then $u_{s+1} = S_0^r q_{m+1} x_1 \cdots x_r \bar{x}_{r+1} S_0^t$ with x_{r+1} in $\{a, b\}$, and $u_{s+2} = S_0^r q_1 x_1 \cdots x_{r+1} S_0^t$. Again, lest $u_1, u_2, \cdots$ terminate, $x_1 \cdots x_{r+1}$ is a word in M. Let τ be the function on $\theta(\Sigma)$ defined by $\tau(\varepsilon) = \varepsilon$, $\tau(\bar{a}) = a$, $\tau(\bar{b}) = b$, $\tau(x) = \varepsilon$ for x in $\Sigma - \{\bar{a}, \bar{b}\}$, and $\tau(y_1 \cdots y_k) = \tau(y_1) \cdots \tau(y_k)$, each y_i in Σ. Then

$$\mathrm{Init}(x_1 x_2 \cdots) = \{\tau(u_1 \cdots u_i)/i \geq 1\}$$

and $x_1 x_2 \cdots$ is a sequence contained in M. Clearly for every sequence $z_1 z_2 \cdots$ contained in M there is a sequence of words $v_1 v_2 \cdots$ in $\theta(\Sigma)$ such that $v_1 = q_3$, $v_i \rightarrow v_{i+1}$ for all i, and $\mathrm{Init}(z_1 z_2 \cdots) = \{\tau(v_1 \cdots v_i)/i \geq 1\}$.

Let c be an element not in Σ. As is easily seen, the set $A_1 = \{u^+ cv/u \rightarrow v\}$ is generated by the grammar $G = (\Sigma \cup \{\xi_1, \cdots, \xi_9, c\}, \Sigma \cup \{c\}, P, \sigma)$, where P contains all the productions (1)–(9) of Lemma 2.4 together with

(10) $\sigma \rightarrow \xi_4$, $\sigma \rightarrow \xi_5$.

(11) $\xi_4 \rightarrow S_0 \xi_4 S_0$, $\xi_4 \rightarrow \xi_6 \bar{a}$, $\xi_4 \rightarrow \xi_6 \bar{b}$.

(12) $\xi_6 \to a\xi_6\, a,\ \xi_6 \to b\xi_6\, b,\ \xi_6 \to q_3\, \xi_7\, q_{m+1}$.
(13) $\xi_7 \to S_0\, \xi_7\, S_0\,,\ \xi_7 \to c$.
(14) $\xi_5 \to S_0\, \xi_5\, S_0\,,\ \xi_5 \to \bar{a}\xi_8\, a,\ \xi_5 \to \bar{b}\xi_8\, b$.
(15) $\xi_8 \to a\xi_8\, a,\ \xi_8 \to b\xi_8\, b,\ \xi_8 \to q_{m+1}\, \xi_9\, q_1$
(16) $\xi_9 \to S_0\, \xi_9\, S_0\,,\ \xi_9 \to c$.

Let $A_2 = \{wcw^+/w \text{ in } \theta(\Sigma)\}$, $A = \mathrm{Init}(q_3\, c(A_1 c)^*)$, and $B = \mathrm{Init}(A_2 c)^*$. A and B are languages whose intersection contains exactly those sequences of the form $v_0\, cv_1\, c \cdots$ with $v_0 = q_3\, cq_3$ and, for all $i \geq 1$, $v_i = u_i\, cu_i^+$ and $u_i \to u_{i+1}$. Since M contains a sequence, so does $A \cap B$. Suppose $A \cap B$ contains a recursive sequence $y_1\, y_2 \cdots$. Then $\tau(y_1)\tau(y_2) \cdots = x_1^2\, x_2^2 \cdots$, each x_i in $\{a, b\}$, is also a recursive sequence. Then $x_1\, x_2 \cdots$ is a recursive sequence in M, a contradiction. Thus $A \cap B$ contains no recursive sequence.

Finally, let τ_1 and τ_2 be the language-preserving functions of Lemma 2.3. Then the language $\tau_1(A) \cup \tau_2(B)$ contains exactly those sequences $y_1^2\, y_2^2 \cdots$ where $y_1\, y_2 \cdots$ is a sequence contained in $A \cap B$. Hence $\tau_1(A) \cup \tau_2(B)$ contains a sequence but no recursive sequence.

THEOREM 3.2. *There exists a language over a two letter alphabet which contains a sequence but no recursive sequence.*

Proof. The theorem follows from Lemma 3.1 in the same manner in which Theorem 2.3 follows from Lemma 2.6.

Remarks. (1) Any language which contains no recursive sequence either contains no sequence or else contains uncountably many sequences. For let L be a language which contains a sequence but no recursive sequence. Suppose there is a word w such that w begins exactly one sequence in L, say the sequence $wy_1\, y_2 \cdots$. Then the language[9] $\{x/wx \text{ in } L\}$ contains the sequence $y_1\, y_2 \cdots$ and no other. By Theorem 3.1, $y_1\, y_2 \cdots$ is then recursive, so that $wy_1\, y_2 \cdots$ is a recursive sequence contained in L, a contradiction. Therefore for every word w which begins a sequence in L there are words w_1 and w_2 such that (i) ww_1 and ww_2 both begin sequences in L; (ii) ww_1 is not an initial subword of ww_2 and ww_2 is not an initial subword of ww_1. But this implies the existence of uncountably many sequences in L.

(2) The method of Lemma 3.1 may be applied to transmit further properties of recursive sets to languages. For example, it can be shown that there exists a recursive set M of words with the property that the set of recursive sequences contained in M is not itself even recursively enumerable. Then the method of Lemma 3.1 allows proof that there exists a language M with the same property.

In passing, we mention two open problems.

(1) Characterize the set of distinguished sequences.

---

[9] It is known that if L is a language and w a word, then $\{x/wx \text{ in } L\}$ is a language [6].

(2) Characterize the set of those sequences D having the property that there exists a language containing D but no u.p. sequence.

BIBLIOGRAPHY

1. Y. Bar-Hillel, M. Perles, and E. Shamir, *On formal properties of simple phase structure grammars*, Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikationsforschung, vol. 14 (1961), pp. 143–172.
2. N. Chomsky, *Three models for the description of language*, IRE Transactions on Information Theory, vol. 2 (1956), pp. 113–124.
3. ———, *On certain formal properties of grammars*, Information and Control, vol. 2 (1959), pp. 137–167.
4. M. Davis, *Computability and unsolvability*, New York, McGraw-Hill, 1958.
5. A. Ehrenfeucht, *Separable theories*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astr. Phys., vol. 9 (1961), pp. 17–19.
6. S. Ginsburg and H. G. Rice, *Two families of languages related to ALGOL*, J. Assoc. Comput. Mach., vol. 9 (1962), pp. 350–371.
7. S. Ginsburg and G. F. Rose, *Operations which preserve definability in languages*, J. Assoc. Comput. Mach., vol. 10 (1963), pp. 175–195.
8. ———, *Some recursively unsolvable problems in ALGOL-like languages*, J. Assoc. Comput. Mach., vol. 10 (1963), pp. 29–47.
9. E. L. Post, *A variant of a recursively unsolvable problem*, Bull. Amer. Math. Soc., vol. 52 (1946), pp. 264–268.
10. M. Rabin and D. Scott, *Finite automata and their decision problems*, IBM Journal of Research and Development, vol. 3 (1959), pp. 114–125.
11. E. Sprecker, *Der Satz vom Maximum in der Rekursiven Analysis*, Constructivity in mathematics, studies in logic and the foundations of mathematics, North Holland Publishing Company, 1959.

System Development Corporation
Santa Monica, California
University of California
Santa Barbara, California