A Note on Undecidable Properties of Formal Languages*

by

SHEILA GREIBACH

Harvard University
Consultant, System Development Corporation

A general set of conditions is given under which a property is undecidable for a family of languages. Examples are given of the application of this result to well-known families of languages.

Recently there have been attempts toward a unified theory of languages and automata by using closure properties to characterize families of languages [4] and abstracting the notion of a class of acceptors [4, 9]. Many of the properties enjoyed by such families as the context-free languages, one-way stack languages, and some complexity classes of languages defined by Turing machines can be established uniformly from machine or language structure. Relations between decision problems for associated classes of machines have also been established [9]. In this paper we establish the undecidability of all properties and all families of languages fulfilling certain specifications. We obtain as corollaries many well-known results such as the undecidability of the inherent ambiguity problem. For example, to establish that it is undecidable whether a context-free language is, say, metalinear, we need only find a context-free language that is not metalinear.

We shall assume the reader to be familiar with the definitions of regular sets and regular expressions [10]. For sets of strings A and B, we use the notation:

$$AB = \{xy \mid x \text{ is in } A, y \text{ is in } B\}$$

$$A/B = \{w \mid \text{there is a } y \text{ in } B \text{ such that } wy \text{ is in } A\}$$

$$B\backslash A = \{w \mid \text{there is a } y \text{ in } B \text{ such that } yw \text{ is in } A\}$$

$$A^+ = AA^*,$$

*Research sponsored in part by the Air Force cambridge Research Laboratories, Office of Aerospace Research, USAF, under Contract F1962867C0008, and by the Air Force Office of Scientific Research, Office of Aerospace Research, USAF under Grant No. AF-AFOSR-1203-67.

MATHEMATICAL SYSTEMS THEORY, Vol. 2, No. 1
Published by Springer-Verlag New York Inc.

where A^* is the monoid freely generated by A with identity ϵ. A language is ϵ-free if it does not contain ϵ; a family of languages is ϵ-free if all its members are ϵ-free. A generalized sequential machine (gsm) is a 6-tuple $M = (K, \Sigma_1, \Delta_1, \delta, \lambda, q_0)$ where $\delta: K \times \Sigma_1 \to K$ and $\lambda: K \times \Sigma_1 \to \Delta_1^*$ are functions. We extend δ and λ to $K \times \Sigma^*$ as follows. If q is in K, x in Σ^*, and a in Σ, then

$$\begin{aligned} \delta(q, xa) &= \delta(\delta(q, x), a) \\ \delta(q, \epsilon) &= q \\ \lambda(q, xa) &= \lambda(q, x)\lambda(\delta(q, x), a) \\ \lambda(q, \epsilon) &= \epsilon. \end{aligned}$$

Then we define

$$M(L) = \{\lambda(q_0, q) \mid w \text{ in } L\}$$
$$M^{-1}(L) = \{y \mid \lambda(q_0, y) \text{ is in } L\}.$$

A homomorphism is a one-state gsm mapping. A function f is ϵ-free if $f(w) = \epsilon$ implies $w = \epsilon$. We let N represent the positive integers.

To say that a property is undecidable for a family of languages is, of course, only an informal way of saying that in a certain enumeration of the family (e.g., by grammars) the set of names of languages with this property is not recursive. We shall introduce the notion of an effective family in order to provide a formal background for our results which will be expressed more informally.

Definition. An effective family of languages is a quintuple $(\Sigma, \mathscr{F}, f_1, f_2, \mu)$ where

(1) Σ is a countable vocabulary and μ a total recursive function such that for any finite subset Σ_1 of Σ, $\mu(\Sigma_1)$ is in $\Sigma - \Sigma_1$. $\mathscr{F}$ is a family of languages.

(2) f_1 is a function from N onto $\mathscr{F}$ such that the mapping g defined on $N \times \Sigma^*$ by

$$g(n, w) = \begin{cases} 1 \text{ if } w \text{ is in } f_1(n) \\ \text{undefined otherwise} \end{cases}$$

is partial recursive.

(3) f_2 is a total recursive function from N into the finite subsets of Σ such that for all n in N,

$$f_1(n) \subseteq [f_2(n)]^* .$$

Notation. Let $\mathscr{R}_\Sigma$ be the regular expressions over Σ and let R in $\mathscr{R}_\Sigma$ define the regular set $\overline{R}$.

Definition. $(\Sigma, \mathscr{F}, f_1, f_2, \mu)$ is effectively closed under a binary operation α on $\mathscr{F}$ if there exists a total recursive function $\bar{\alpha}$: $N \times N \to N$ such that

$f_1(\bar{\alpha}(n_1, n_2)) = \alpha(f_1(n_1), f_1(n_2))$. $(\Sigma, \mathscr{F}, f_1, f_2, \mu)$ is effectively closed under a binary operation β on $\mathscr{F}$ and the regular sets if there exists a total recursive function $\bar{\beta}$ on $N \times \mathscr{R}_\Sigma$ such that $f_1(\beta(n, R)) = \bar{\beta}(f_1(n), \bar{R})$.

Definition. A property P is *undecidable* for $(\Sigma, \mathscr{F}, f_1, f_2, \mu)$ if P is defined on members of $\mathscr{F}$ and the set

$$\{n \mid f_1(n) \text{ has } P\}$$

is not recursive.

Observe that a property undecidable for $(\Sigma, \mathscr{F}, f_1, f_2, \mu)$ might be decidable for $(\Sigma, \mathscr{F}, f_1', f_2, \mu)$ if $f_1 \neq f_1'$, since f_1' might not be effectively constructible from f_1 or vice versa.

In the rest of the paper we shall assume that Σ, f_1, f_2, and μ are fixed and use $\mathscr{F}$ instead of $(\Sigma, \mathscr{F}, f_1, f_2, \mu)$ and write $L \subseteq \Sigma_1^*$ instead of $f_1(n) \subseteq [f_2(n)]^*$. Note that we demand only that $[f_2(n)]^*$ be some finitely generated monoid containing $f_1(n)$; there may be $\Sigma_1 \subset f_2(n)$ with $f_1(n)$ also contained in Σ_1^*. Thus, most finite specifications of languages (such as grammars, acceptors, regular expressions, etc.) include a maximal vocabulary but usually not a minimal necessary vocabulary.

The main results of this paper appear in two theorems. Each theorem is followed by corollaries which give examples of the use of the theorem to obtain generally known results in language theory. For the purpose of exhibiting these examples, we assume the reader to be familiar with the theory of context-free and context-sensitive languages; most of the background material can be found in reference [2].

THEOREM 1. *Let $\mathscr{F}$ be effectively closed under union and under concatenation by regular sets and let "$L_1 = \Sigma_1^*$" be undecidable for L_1 in $\mathscr{F}$.*[1] *If P is any property that is defined on $\mathscr{F}$ and*

(a) is false for at least one L_2 in $\mathscr{F}$,

(b) is true for all regular sets,

(c) is preserved by inverse gsm, union with $\{\epsilon\}$, and intersection with regular sets,

then P is undecidable for $\mathscr{F}$.

Proof. Let property P be false for L_2 in $\mathscr{F}$, $L_2 \subseteq \Sigma_2^*$. Let L_1 be in $\mathscr{F}$ with $L_1 \subseteq \Sigma_1^*$. Let c be a new symbol. (In our formalism, $L_1 = f_1(n_1)$ and $L_2 = f_1(n_2)$ for some n_1 and n_2 in N. Then $\Sigma_1 = f_2(n_1)$ and $\Sigma_2 = f_2(n_2)$ and $c = \mu(f_2(n_1) \cup f_2(n_2))$, so that c is in $\Sigma - (\Sigma_1 \cup \Sigma_2)$.) Then $L = L_1 c \Sigma_2^* \cup \Sigma_1^* c L_2$ is in $\mathscr{F}$. (Note that L is effectively in $\mathscr{F}$. That is, given n_1 and n_2, we can effectively find m in N such that $f_1(m) = L$.) We claim that P is true for L if and only if $L_1 = \Sigma_1^*$. For if $L_1 = \Sigma_1^*$, then $L = \Sigma_1^* c \Sigma_2^*$ is regular and hence has property P. Otherwise, let y be in $\Sigma_1^* - L_1$. Then, if L has P, so does

$$L \cap y c \Sigma_2^* = y c L_2 \,.$$

[1] More formally, "$f_1(n) = (f_2(n))^*$" is undecidable for $(\Sigma, \mathscr{F}, f_1, f_2, \mu)$.

Let M be the gsm

$$M = (\{q_0, q_1\}, \Sigma_2, \Sigma_1 \cup \Sigma_2 \cup \{c\}, \delta, \lambda, q_0) ,$$

where, for all a in Σ_2,

$$\delta(q_0, a) = \delta(q_1, a) = q_1$$

$$\lambda(q_0, a) = yca$$

$$\lambda(q_1, a) = a.$$

Then, if L has P, so does $M^{-1}(L \cap yc\Sigma_2^*) = L_2 - \{\epsilon\}$, and hence so does L_2. But L_2 does not have P. Hence L has P if and only if $L_1 = \Sigma_1^*$, which is undecidable. Hence, if

$$\{n \mid f_1(n) \text{ has property } P\}$$

is recursive, so is

$$\{n \mid f_1(n) = [f_2(n)]^*\} ,$$

contrary to hypothesis.

Remark. We could replace conditions (a), (b), and (c) by the following.

(a) There is an ϵ-free language L_2 in $\mathscr{F}$ which does not have property P.

(b) All sets of the form $\Sigma_1^* c \Sigma_2^*$ have property P.

(c) If L has P, R is regular and y is a string, then $L \cap R$ and $y \backslash L = \{w \mid yw \text{ is in } L\}$ have property P.

If we define an effective AFL as an effective family of languages effectively closed under union, concatenation, +, ϵ-free homomorphism, intersection with regular and inverse homomorphisms, we obtain the following corollary.

COROLLARY. *If $\mathscr{F}$ is an effective AFL, if "$L = \Sigma_1^*$" is undecidable for L in $\mathscr{F}$, and if $\mathscr{F}_1$ is any proper subfamily closed under union with $\{\epsilon\}$, inverse gsm and intersection with regular homomorphisms, then "L is in $\mathscr{F}_1$" is undecidable.*

The next two corollaries depend on elementary properties of context-free languages and regular sets, and give results that were established in [1, 3, 7, 8].

COROLLARY. *It is undecidable whether a context-free language is (a) regular, (b) linear context-free, (c) deterministic context-free, (d) inherently ambiguous.*

COROLLARY. *It is undecidable whether the complement of a context-free language is (a) regular, (b) linear context-free, (c) context-free.*

COROLLARY. *If $\mathscr{F}$ is any AFL properly contained in the context-free languages, it is undecidable whether a context-free language belongs to $\mathscr{F}$.*

COROLLARY. *It is undecidable whether a context-free language L has the property that $L \cap L_1$ is context-free for all context-free languages L_1.*

Proof. The regular sets have this property [1], as do all context-free subsets of a^*b^* [6]. An ϵ-free language without this property is

$\{a^n b^n c^m \mid n, m \geq 1\}$. The property is obviously preserved under intersection with regular sets but not by inverse gsm. But, if $L \subseteq \Sigma_1^*$ and y is in Σ_1^*, then

$$(y \backslash L) \cap L_1 = [y \backslash (L \cap y\Sigma_1^*)] \cap L_1 = y \backslash (L \cap y\Sigma_1^* \cap yL_1) ,$$

which is context-free if L and L_1 are context-free, and L has the desired property.

The next corollary appears as a theorem in [5].

COROLLARY. *It is recursively unsolvable to determine whether an arbitrary one-way stack language is (a) regular, (b) context free, (c) a deterministic one-way stack language.*

THEOREM 2. *Let $\mathscr{F}$ be effectively closed under concatenation. Let "$L_1 = \varnothing$" be undecidable for $\mathscr{F}$. If P is any property which (a) is false for some ϵ-free L_2 in $\mathscr{F}$, (b) is true of $\varnothing$, and (c) is preserved by inverse gsm and intersection with regular sets, then P is undecidable for $\mathscr{F}$.*

Proof. Let $L_1 \subseteq \Sigma_1^*$, let $L_2 \subseteq \Sigma_2\Sigma_2^*$, and let c be new. Let $L = L_1 c L_2$. Repeating the arguments of the proof of Theorem 1, we see that L has property P if and only if $L_1 = \varnothing$, which is undecidable.

COROLLARY. *It is undecidable whether a context-sensitive language is context-free.*

Definition. If $\mathscr{F}_1$ and $\mathscr{F}_2$ are families of languages, let $\mathscr{F}_1 \wedge \mathscr{F}_2$ be the family of all languages $L_1 \cap L_2$ where L_1 is in $\mathscr{F}_1$, and L_2 is in $\mathscr{F}_2$.

COROLLARY. *If $\mathscr{F}_1$ and $\mathscr{F}_2$ are effectively closed under concatenation, if "$L_1 \cap L_2 = \varnothing$" is undecidable for L_1 in $\mathscr{F}_1$ and L_2 in $\mathscr{F}_2$, and if P is any property which (a) is false for some ϵ-free language $L_3 \cap L_4$, where L_3 is in $\mathscr{F}_1$ and L_4 is in $\mathscr{F}_2$, and (b) is true of the empty set, and (c) is preserved by inverse gsm and intersection with regular sets, then P is undecidable for $\mathscr{F}_1 \wedge \mathscr{F}_2$.*

Proof. Notice that if L_1 and L_3 are in $\mathscr{F}_1$, L_2 and L_4 are in $\mathscr{F}_2$, and c is new, then

$$L = (L_1 \cap L_2)c(L_3 \cap L_4) = L_1 c L_3 \cap L_2 c L_4 ,$$

and so L is in $\mathscr{F}_1 \wedge \mathscr{F}_2$. Again, P is true of L if and only if $L_1 \cap L_2 = \varnothing$.

COROLLARY. *For context-free languages L_1 and L_2, it is undecidable whether $L_1 \cap L_2$ is (a) regular, (b) linear context-free, (c) context-free.*

REFERENCES

[1] Y. Bar-Hillel, M. Perles and E. Shamir, On formal properties of simple phrase structure grammars. *Z. Phonetik Sprachwiss. Kommunikat.* **14** (1961), 143–172.

[2] S. Ginsburg, *The Mathematical Theory of Context-Free Languages.* McGraw-Hill Book Company, New York, 1966.

[3] S. Ginsburg and S. Greibach, Deterministic context-free languages. *Information and Control* **9** (1966), 620–648.

[4] S. Ginsburg and S. Greibach, Abstract families of languages. SDC Document TM-738/031/00, April 1967.

[5] S. GINSBURG, S. GREIBACH and M. HARRISON, One-way stack automata. *J. Assoc. Comput. Mach.* **14** (1967), 389–418.
[6] S. GINSBURG and E. H. SPANIER, Bounded ALGOL-like languages. *Trans. Amer. Math. Soc.* **113** (1964), 333–368.
[7] S. GINSBURG and J. S. ULLIAN, Ambiguity in context-free languages. *J. Assoc. Comput. Mach.* **13** (1966), 62–89.
[8] S. GREIBACH, The unsolvability of the recognition of linear context-free languages. *J. Assoc. Comput. Mach.* **13** (1966), 582–587.
[9] J. HOPCROFT and J. ULLMAN, An approach to a unified theory of automata. (To appear in *Bell System Tech. J.*)
[10] S. RABIN and D. SCOTT, Finite automata and their decision problems. *IBM J. Res. Develop.* **3** (1959), 114–125.

(Received 29 September 1967)