TM-738/038/00

A Note on Undecidable Properties of Formal Languages

SCIENTIFIC REPORT NO. 11

Sheila Greibach 28 August 1967

Contract Monitor Thomas V. Griffiths
Data Sciences Labortory

Monitored by: R. Swanson, SRIR (AFOSR) Supported by
SDC's Independent Research Program;
Contract F1962867C0008, Programming (Algorithmic)
Languages, Project No. 5632,
Task No. 563205, and Grant No. AF AFOSR 1203 67
Work Unit No. 56320501

Prepared for
AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH, USAF
BEDFORD, MASSACHUSETTS
and
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
OFFICE OF AEROSPACE RESEARCH, USAF
ARLINGTON, VIRGINIA

TM-738/038/00

AFCRL - 67-0539

SDC TM SERIES

The work reported herein was supported by SDC and Contract F1962867C0008, Programming (Algorithmic) Languages; Project No. 5632, Task No. 563205; Work Unit No. 56320501; and Grant No. AF-AFOSR-1203-67.

SCIENTIFIC REPORT NO. 11

A Note on Undecidable Properties of Formal Languages

by
Sheila Greibach*

28 August 1967

Monitored by: Contract Monitor
Thomas V. Griffiths
Data Sciences Laboratory
R. Swanson, SRIR (AFOSR)

Prepared for
AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH, USAF
BEDFORD, MASSACHUSETTS
and
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
OFFICE OF AEROSPACE RESEARCH, USAF
ARLINGTON, VIRGINIA

SYSTEM
DEVELOPMENT
CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA
90406

*Harvard University and
Consultant, System Development Corporation

ABSTRACT

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.

A NOTE ON UNDECIDABLE PROPERTIES OF FORMAL 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:

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

$$A/B = \{w / \exists\, y \text{ in } B \text{ such that } wy \text{ is in } A\}$$

$$B\backslash A = \{w / \exists\, y \text{ in } B \text{ such that } yw \text{ is in } A\}$$

$$A^+ = AA^*$$

*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 AFOSR Grant No. AF-AFOSR-1203-67.

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 $M = (K, \Sigma_1, \Delta_1, \delta, \lambda, q_o)$ 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

$$\delta(q,xa) = \delta(\delta(q,x),a)$$

$$\delta(q,\epsilon) = q$$

$$\lambda(q,xa) = \lambda(q,x)\lambda(\delta(q,x),a)$$

$$\lambda(q,\epsilon) = \epsilon.$$

Then we define

$$M(L) = \{\lambda(q_o,q)/w \text{ in } L\}$$

$$M^{-1}(L) = \{y/\lambda(q_o,y) \text{ is in } L\}.$$

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.

<u>Definition</u>. An effective family of languages is a quintuple $(\Sigma, \mathcal{F}, f_1, f_2, \mu)$ 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 $\Sigma-\Sigma_1$. $\mathcal{F}$ is a family of languages.

(2) f_1 is a function from N onto $\mathcal{F}$ 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,

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

<u>Notation</u>. Let $\mathcal{R}_\Sigma$ be the reguler expressions over Σ and let R in $\mathcal{R}_\Sigma$ define the regular set $\overline{R}$.

<u>Definition</u>. $(\Sigma,\mathcal{F},f_1,f_2,\mu)$ is effectively closed under a binary operation α on $\mathcal{F}$ if there exists a total recursive function $\overline{\alpha} : N \times N \rightarrow N$ such that $f_1(\overline{\alpha}(n_1,n_2)) = \alpha(f_1(n_1),f_1(n_2))$. $(\Sigma,\mathcal{F},f_1,f_2,\mu)$ is effectively closed under a binary operation β on $\mathcal{F}$ and the regular sets if there exists a total recursive function $\overline{\beta}$ on $N \times \mathcal{R}_\Sigma$ such that $f_1(\overline{\beta}(n,R)) = \beta(f_1(n),\overline{R})$.

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

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

is not recursive.

Observe that a property undecidable for $(\Sigma,\mathcal{F},f_1,f_2,\mu)$ might be decidable for $(\Sigma,\mathcal{F},f_1',f_2,\mu)$ 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 $\mathcal{F}$ instead of $(\Sigma, \mathcal{F}, f_1, f_2, \mu)$ 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 $\mathcal{F}$ be effectively closed under union and under concatenation by regular sets and let "$L_1 = \Sigma_1^*$" be undecidable for L_1 in $\mathcal{F}$.(1) If P is any property that is defined on $\mathcal{F}$ and

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

(b) is true for all regular sets and

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

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

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

<u>Proof</u>. Let property P be false for L_2 in $\mathcal{F}$, $L_2 \subseteq \Sigma_2^*$. Let L_1 be in $\mathcal{F}$ with $L_1 \subseteq \Sigma_1^*$. Let c be a new symbol.[(2)] Then $L = L_1 c \Sigma_2^* \cup \Sigma_1^* c L_2$ is in $\mathcal{F}$.[(3)] 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

$$L \cap yc\Sigma_2^* = ycL_2.$$

Let M be the gsm

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

where

$$\left.\begin{array}{l} \delta(q_0, a) = \delta(q_1, a) = q_1 \\ \lambda(q_0, a) = yca \\ \lambda(q_1, a) = a \end{array}\right\} \quad \text{all } a \text{ in } \Sigma_2.$$

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

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

is recursive, so is

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

contrary to hypothesis.

---

(2) 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 $\Sigma - (\Sigma_1 \cup \Sigma_2)$.

(3) Note that L is effectively in $\mathcal{F}$. That is, given n_1 and n_2, we can effectively find m in N such that $f_1(m) = L$.

<u>Remark</u>. We could replace conditions (a), (b), and (c) by:

(a) There is an ϵ-free language L_2 in $\mathcal{F}$ 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 $y \backslash L = \{w / yw \text{ is in } L\}$

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 homomorphism, we obtain as a corollary:

<u>Corollary</u>. If $\mathcal{F}$ is an effective AFL, if "$L = \Sigma_1^*$" is undecidable for L in $\mathcal{F}$, and if $\mathcal{F}_1$ is any proper subfamily closed under union with $\{\epsilon\}$, inverse gsm and intersection with regular, then "L is in $\mathcal{F}_1$" 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].

<u>Corollary</u>. It is undecidable whether a context-free language is:

(a) regular.

(b) linear context-free.

(c) deterministic context-free.

(d) inherently ambiguous.

<u>Corollary</u>. It is undecidable whether the complement of a context-free language is:

(a) regular.

(b) linear context-free.

(c) context-free.

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

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 $\{a^n b^n c^m / n, m \geq 1\}$. 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 $\mathcal{F}$ be effectively closed under concatenation. Let "$L_1 = \emptyset$" be undecidable for $\mathcal{F}$. If P is any property which

(a) is false for some ϵ-free L_2 in $\mathcal{F}$,

(b) is true of $\emptyset$, and

(c) is preserved by inverse gsm and intersection with regular,

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

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 $L_1 = \emptyset$, which is undecidable.

Corollary. It is undecidable whether a context-sensitive language is context free.

Definition. If $\mathcal{F}_1$ and $\mathcal{F}_2$ are families of languages, let $\mathcal{F}_1 \wedge \mathcal{F}_2$ be the family of all languages

$$L_1 \cap L_2$$

where L_1 is in $\mathcal{F}_1$, and L_2 is in $\mathcal{F}_2$.

Corollary. If $\mathcal{F}_1$ and $\mathcal{F}_2$ are effectively closed under concatenation, if "$L_1 \cap L_2 = \emptyset$" is undecidable for L_1 in $\mathcal{F}_1$ and L_2 in $\mathcal{F}_2$, 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 $\mathcal{F}$, and L_4 in $\mathcal{F}_2$, 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 $\mathcal{F}_1 \wedge \mathcal{F}_2$.

Proof. Notice that if L_1 and L_3 are in $\mathcal{F}_1$ and L_2 and L_4 are in $\mathcal{F}_2$ 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 $\mathcal{F}_1 \wedge \mathcal{F}_2$. Again, P is true of L if and only if $L_1 \cap L_2 = \emptyset$.

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.

BIBLIOGRAPHY

[1] Bar-Hillel, Y., Perles, M., and Shamir, E. "On formal properties of simple phrase structure grammars." Z. Phonetik, Sprachwissen. Kommunikationsforsch, Vol. 14, 1961, pp. 143-172.

[2] Ginsburg, S. The mathematical theory of context-free languages. McGraw-Hill Book Company, New York, 1966.

[3] Ginsburg, S. and Greibach, S. "Deterministic context-free languages." Information and Control, Vol. 9, 1966, pp. 620-648.

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

[5] Ginsburg, S., Greibach, S., and Harrison, M. "One-way stack automata." Journal of the Association for Computing Machinery, Vol. 14, 1967, pp. 389-418.

[6] Ginsburg, S. and Spanier, E. H. "Bounded ALGOL-like languages." Transactions of the American Mathematical Society, Vol. 113, 1964, pp. 333-368.

[7] Ginsburg, S. and Ullian, J. S. "Ambiguity in context-free languages." Journal of the Association for Computing Machinery, Vol. 13, 1966, pp. 62-89.

[8] Greibach, S. "The unsolvability of the recognition of linear context-free languages." Journal of the Association for Computing Machinery, Vol. 13, 1966, pp. 582-587.

[9] Hopcroft, J. and Ullman, J. "An approach to a unified theory of automata." (To appear in Bell System Technical Journal.)

[10] Rabin, S. and Scott, D. "Finite automata and their decision problems." IBM Research Development Journal, Vol. 3, 1959, pp. 114-125.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) System Development Corporation 2500 Colorado Avenue Santa Monica, California 90406	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP
3. REPORT TITLE A NOTE ON UNDECIDABLE PROPERTIES OF FORMAL LANGUAGES	
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Scientific. Interim.	
5. AUTHOR(S) (Last name, first name, initial) Sheila Greibach	

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
28 August 1967	11	10

8a. CONTRACT OR GRANT NO. Contract F19628-67-C-0008 and Grant No. AF-AFOSR-1203-67 b. Project, Task, Work Unit Nos. 5632- 05- 01 c. DoD Element 61445014 d. DoD Subelement 681305	9a. ORIGINATOR'S REPORT NUMBER(S) TM-738/038/00 Scientific Report No. 11 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AFCRL-67-0539
10. AVAILABILITY/LIMITATION NOTICES Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.	
11. SUPPLEMENTARY NOTES prepared for: Hq. AFOSR, OAR (SIR) United States Air Force, Arlington, Virginia	12. SPONSORING MILITARY ACTIVITY Air Force Cambirdge Research Laboratories (CRB) L. G. Hanscom Field Bedford, Massachusetts 01730

13. ABSTRACT

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.

DD FORM 1 JAN 64 1473

Unclassified

Security Classification

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Formal languages Undecidable properties Families of Languages Automata Linguistics						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through ________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through ________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through ________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Unclassified

Security Classification