DISCRETE MATHEMATICS 2 (1972) 347–355. North-Holland Publishing Company

A GENERALIZATION OF PARIKH'S SEMILINEAR THEOREM

S. A. GREIBACH*

Department of System Science, University of California, Los Angeles, Calif. 90024, U.S.A.

Received 12 November 1971**

Abstract. The operation of nested iterated substitution preserves languages with the semilinear property. As a consequence, the following generalization of Parikh's theorem that each context-free language has the semilinear property is obtained: If $\mathcal{L}$ is a family of languages with the semilinear property closed under intersection with regular sets and finite substitution, all members of the least superAFL containing $\mathcal{L}$ have the semilinear property.

It was first proven by Parikh (in 1960) that counting the occurrences of symbols in words of a context-free language yields a semilinear set [5]. Subsequent proofs of this theorem have used variants of the original approach utilizing properties of derivation trees in context-free grammars. In this paper we use the study of operations on languages to prove a generalization of the Parikh theorem and obtain, as a corollary, an independent proof of the original result.

Families of languages having the semilinear property (called slip languages) were extensively studied in [1] and various closure properties were established. The operation of nested iterated substitution was introduced in [2]. We now show that nested iterated substitution preserves slip languages. This shows that if L is a nontrivial family of slip languages closed under intersection with regular sets, and finite substitution, then the least superAFL containing L is a slip family. Since the context-free languages are the least superAFL containing the finite sets, this shows that the context-free languages form a slip AFL.

First we must repeat some definitions regarding families of languages,

* The research represented in this paper was supported in part by the National Science Foundation under Grant No. GJ-803.

** Original version received 21 April 1971; revised version received 5 July 1971.

operations and semilinear sets. In this paper we assume that we have a fixed infinite set Σ of symbols; by a "symbol" we mean a member of Σ and by a "vocabulary" we mean a finite subset of Σ.

Definition 1. A family $\mathcal{L}$ is a *nontrivial* family of languages if $\mathcal{L}$ contains at least one nonempty language and if for each L in $\mathcal{L}$ there is a vocabulary Σ_1 such that $L \subseteq \Sigma_1^*$. A *full AFL (abstract family of languages)* is a nontrivial family of languages closed under union, concatenation, Kleene *, intersection with regular sets, homomorphism and inverse homomorphism.[‡]

For a family of languages $\mathcal{L}$, we let $\hat{F}(\mathcal{L})$ be the least full AFL containing $\mathcal{L}$. We write $\hat{F}(L)$ for $\hat{F}(\{L\})$.

Definition 2. Let $L \subseteq \Sigma_1^*$ and for each a in Σ_1, let $L_a \subseteq \Sigma_a^*$, where Σ_1 and Σ_a are vocabularies. Let $\tau(a) = L_a$ for each a in Σ_1. Let $\tau(xy) = \tau(x)\tau(y)$ and $\tau(L) = \bigcup_{w \in L} \tau(w)$. Then τ is a *substitution*. If $\mathcal{L}$ contains $\tau(L)$ whenever $L \in \mathcal{L}$ and $\tau(a) \in \mathcal{L}$ for each a, then $\mathcal{L}$ is *substitution closed*. If $\tau(a)$ is finite for each a, then τ is a *finite substitution*.

Definition 3. Let $L \subseteq \Sigma_1^*$ and let τ be a substitution, with $\tau(a) \subseteq \Sigma_a^*$ for each a in Σ_1. Extend τ to $\Sigma_1 \cup (\bigcup_a \Sigma_a)$ by defining $\tau(b) = \{b\}$ for $b \in (\bigcup_a \Sigma_a) - \Sigma_1$. Let $\tau^1(L) = \tau(L)$, and $\tau^{n+1}(L) = \tau(\tau^n(L))$ for $n \geq 1$. Let $\tau^*(L) = \bigcup_n \tau^n(L)$. Then τ^* is an *iterated substitution*. It is a *nested iterated substitution* (abbreviated n.i.s.) if $a \in \tau(a)$ for all a in Σ_1. It is a *simple n.i.s.* if $\tau(a) \neq \{a\}$ for at most one a in Σ_1. A family of languages $\mathcal{L}$ is *closed under nested iterated substitution* if $\tau^*(L)$ is in $\mathcal{L}$ whenever L is in $\mathcal{L}$, τ^* is a nested iterated substitution and $\tau(a)$ is in $\mathcal{L}$ for all $a \in \Sigma_1$.

The operation of simple n.i.s. was investigated in [3] and [4]. Later, we reduce any nested iterated substitution to a series of simple ones.

Definition 4. A *superAFL* is a full AFL closed under nested iterated substitution.

[‡] For a set A we let A^* be the monoid generated by A with identity e.

For a family of languages L, we let $\hat{S}^{\infty}(L)$ be the least superAFL containing L. In [2], a superAFL was defined using fewer operations and shown to be equivalent to Definition 4, which is more convenient for our present purposes. We write $\hat{S}^{\infty}(\mathrm{L})$ for $\hat{S}^{\infty}(\{\mathrm{L}\})$.

We let N be the set of nonnegative integers and let N^n denote all n-dimensional vectors over N; when dealing with vectors, we use the usual notation for vector addition and subtraction and scalar multiplication.

For subsets C and P of N^n we let $Q(C;P)$ denote the least subset of N^n containing C and closed under addition by members of P. If $C = \{\mathrm{c}\}$, we write $Q(\mathrm{c};P)$ for $Q(\{\mathrm{c}\};P)$. Members of C are called *constants* and members of P, *periods*.

Definition 5. A subset Q of N^n is *linear* if $Q = Q(\mathrm{c};P)$ for some c in N^n and some finite subset P of N^n. A subset of N^n is *semilinear* if it is the finite union of linear sets.

When dealing with a vocabulary Σ_1 we assume that we are supplied with an ordering $a_1, \ldots, a_n$ of the elements of Σ_1. It is immaterial which ordering is given, so we may select one to suit our convenience. We define a homomorphism ν from the monoid Σ_1^* under concatenation to the monoid N^n under vector addition, by $\nu(a_i) = (t_{i1}, \ldots, t_{in})$, where $t_{ij} = 0$ for $i \neq j$ and $t_{ii} = 1$. We call such a homomorphism a *Parikh mapping*.

Definition 6. A language $\mathrm{L} \subseteq \Sigma_1^*$ is a *language with the semilinear property* (abbreviated *slip* language) if $\nu(\mathrm{L})$ is semilinear for a Parikh mapping, ν, of Σ_1^*. If $\nu(\mathrm{L})$ is linear, L has the *linear property*. A slip family of languages is a family containing only slip languages.

We need the following lemma from [1]:

Lemma 1. *If L is a nontrivial slip family closed under intersection with regular sets and finite substitution, then $\hat{F}(\mathrm{L})$ is a slip family.*

Our main result, Theorem 2 below, extends Lemma 1 to show that under those hypotheses, $\hat{S}^{\infty}(L)$ is also a slip family. To do this we prove that nested iterated substitution preserves slip languages. First, we need closure under substitution. This was shown in [1], but the proof used

the fact that regular languages form a slip family. Since we want an independent proof of the Parikh theorem and its ramifications, we need an independent proof of the fact that regular languages are slip languages. This is provided by the next lemma.

Lemma 2. *The family of slip languages is closed under union, concatenation, Kleene *, and intersection with regular sets of the forms* T^* *and* T^*aT^*, *where* T *is a vocabulary and a is a symbol.*

Proof. Closure under union is trivial. Since concatenation distributes over union and we have closure under union, we need consider only the concatenation of languages with the linear property. If $\Sigma_1 \subseteq \Sigma_2$, $L \subseteq \Sigma_1^*$, and ν and μ are Parikh mappings of Σ_1^* and Σ_2^* respectively, then $\nu(L)$ is semilinear if and only if $\mu(L)$ is semilinear. So we shall henceforth let all languages have the same vocabulary. If $L_1 \cup L_2 \subseteq \Sigma_1^*$, $\nu(L_1) = Q(c_1; P_1)$ and $\nu(L_2) = Q(c_2; P_2)$ for Parikh mapping ν, obviously $\nu(L_1 L_2) = Q(c_1 + c_2; P_1 \cup P_2)$; so concatenation preserves languages with the linear property as well as slip languages.

Let $L \subseteq \Sigma_1^*$ and $\nu(L) = Q(c_1; P_1) \cup \ldots \cup Q(c_n; P_n)$. For each subset $S \subseteq \{1, \ldots, n\}$, e.g., $S = \{i_1, \ldots, i_m\}$, $i_1 < \ldots < i_m$, let $c_S = c_{i_1} + c_{i_2} + \ldots + c_{i_m}$ and $P_S = \bigcup_{i_k \in S} (P_{i_k} \cup \{c_{i_k}\})$. Then

$$\nu(L^*) = \bigcup_{S \subseteq \{1,\ldots,n\}} Q(c_S; P_S) \cup Q(0; \{0\}),$$

where **0** is the appropriate vector with all coordinates zero.

If $L \subseteq \Sigma_1^*$ is a slip language, $a \in T \subseteq \Sigma_1$, and ν is a Parikh mapping of Σ_1^*, obviously $\nu(L \cap T^*)$ is obtained from a semilinear representation of $\nu(L)$ by omitting any linear component whose constant has a nonzero coordinate corresponding to any member of $\Sigma_1 - T$, and then omitting from any remaining linear component any period with a nonzero coordinate corresponding to some member of $\Sigma_1 - T$. Similarly, a semilinear representation of $\nu(L \cap T^*aT^*)$ is obtained from one for $\nu(L \cap T^*)$ by replacing any linear component $Q(c; P)$, where the a-coordinate of c is zero by the (finite) union of all $Q(c + p; P)$, where p is in P and the a-coordinate of p is nonzero.

Corollary 1. *Every regular language is a slip language.*

In general, intersection with regular languages does not preserve slip languages. We shall need the particular closure results established above.

In [1] it was shown that if L is a slip language, τ a substitution by slip languages and ν an appropriate Parikh mapping, then $\nu(\tau(L)) = \nu(R)$ for some regular set R. Combining this with the corollary above proves closure under substitution without appeal to the Parikh theorem:

Lemma 3. *The family of slip languages is closed under substitution.*

We need one more lemma establishing closure under simple n.i.s.; this is our main technical lemma.

Lemma 4. *Let τ be a simple n.i.s. defined on Σ_1 such that $\tau(b)$ is a slip language for all $b \in \Sigma_1$. Then for $b \in \Sigma_1$, $\tau^\infty(\{b\})$ is a slip language.*

Proof. Let $a \in \Sigma_1$ and $\tau(b) = \{b\}$ for $b \neq a$. Obviously $\tau^\infty(\{b\}) = \{b\}$ is a slip language for $b \neq a$, so we need only consider $\tau^\infty(\{a\})$.

By Lemma 2, $\tau(a) \cap (\Sigma_1 - \{a\})^*$ and $\tau(a) \cap \Sigma_1^* a \Sigma_1^*$ are slip languages. Hence, if $\tau_1(a) = [\tau(a) \cap (\Sigma_1 - \{a\})^*] \cup \{a\}$, $\tau_2(a) =$, $\tau(a) \cap \Sigma_1^* a \Sigma_1^*$ and $\tau_1(b) = \tau_2(b) = \{b\}$ for $b \neq a$, then τ_1 and τ_2 are substitutions by slip languages and τ_2^∞ is a simple n.i.s. Clearly $\tau^\infty(\{a\}) = \tau_1(\tau_2^\infty(\{a\}))$ since substitution by a word in $\tau_1(a)$ precludes further iteration. Hence, in view of Lemma 3, it suffices to consider only the case $\tau = \tau_2$; i.e., $\tau(a) \subseteq \Sigma_1^* a \Sigma_1^*$.

Let $\Sigma_1 = \{a_1, \ldots, a_n\}$, with Σ_1 ordered so that $a = a_1$, and let ν be the corresponding Parikh mapping. Let $\nu(\tau(a)) = Q(c_1; P_1) \cup \ldots \cup Q(c_m; P_m)$, $c_i \in N^n$, and P_i finite for $1 \leqslant i \leqslant m$. Let $1 = (1, 0, \ldots, 0)$ and $0 = (0, 0, \ldots, 0)$. Because $\tau(a) \subseteq \Sigma_1^* a \Sigma_1^*$, the first coordinate of each c_i is nonzero and $c_i - 1 \in N^n$.

For each nonempty subset S of $\{1, \ldots, m\}$, e.g., $S = \{i_1, \ldots, i_t\}$, $i_1 < \ldots < i_t$, let $c_S = c_{i_1} + \ldots + c_{i_t} - (t-1)\,1$ and $P_S = \bigcup_{i_j \in S} (P_{i_j} \cup \{c_{i_j} - 1\})$. Since $c_{i_j} - 1 \in N^n$, $c_S \in N^n$ and P_S is a finite subset of N^n.

Let

$$L = \bigcup_{\substack{S \subseteq \{1,\ldots,m\} \\ S \neq \emptyset}} Q(c_S; P_S).$$

We claim that $L = \nu(\tau^*(\{a\}))$.

Intuitively, component $Q(c_S; P_S)$ corresponds to those words obtained by substitutions involving *all* and *only* components $Q(c_i; P_i)$ with $i \in S$. The order in which substitutions are performed or where they occur does not affect the value of $\nu(w)$; also there is no difference between substitution of w_1 and w_2 with $\nu(w_1) = c + x$ and $\nu(w_2) = c + y$ or substituting w_3 and w_4 with $\nu(w_3) = c + x + y$ and $\nu(w_4) = c$.

Since $a \in \tau(a)$, we can consider each application of τ to involve replacing exactly one occurrence of a with a word in $\tau(a)$. To show that $\nu(\tau^*(\{a\})) \subseteq L$, we proceed by induction on the number of applications of τ. Clearly

$$\nu(\tau(\{a\})) = \bigcup_i Q(c_i; P_i) \subseteq \bigcup_i L_{\{i\}} \subseteq L.$$

Suppose we have established the result for $l \geq 1$ applications and w is obtained by $l + 1$ applications of τ. Then $w = xyz$, where xaz is obtained in l applications of τ and $y \in \tau(a)$. By hypothesis, $\nu(y) = c_i + u$ for some i, with $u \in Q(0; P_i)$. By the induction hypothesis, $\nu(xaz) = c_S + v$ for some $S \subseteq \{1, \ldots, m\}$ and $v \in Q(0; P_S)$. Hence

$$\nu(w) = \nu(xaz) + \nu(y) - 1 = c_S + v + c_i + u - 1$$

$$= c_S + (v + u) + (c_i - 1).$$

If $i \in S$, u and $c_i - 1$ are in $Q(0; P_S)$, so $\nu(w) \in Q(c_S; P_S) \subseteq L$. If $i \notin S$, let $T = S \cup \{i\}$. Then $c_T = c_S + c_i - 1$, $P_S \subseteq P_T$ and $P_i \subseteq P_T$, so

$$\nu(w) = c_T + (v + u) \in Q(c_T; P_T) \subseteq L.$$

The proof that $L \subseteq \nu(\tau^*(\{a\}))$ also depends on the fact that $a \in \tau(a)$. We proceed by induction on $\#S$. If $S = \{i\}$ and $\alpha \in L_S$, then $\alpha = c_i + u + t(c_i - 1)$ for $u \in Q(0; P_i)$ and $t \geq 0$. Since $c_i + u$ and c_i are in $Q(c_i, P_i) \subseteq \nu(\tau(a))$, there are xay and waz in $\tau(a)$ with $\nu(xay) = c_i + u$ and $\nu(waz) = c_i$. Then

$$\bar{w} = xw^t az^t y \in \tau^*(\{a\})$$

and

$$\nu(\bar{w}) = \nu(xay) + t\,\nu(waz) - t(1) = a.$$

Now assume we have shown $Q(c_S; P_S) \subseteq \nu(\tau^*(\{a\}))$ for $\emptyset \neq S' \neq S$. Let $S = T \cup \{t\}$, $t \notin T$. If $a \in Q(c_S; P_S)$, then $a = c_T + (c_t - 1) + u + v$, $u \in Q(0; P_T)$ and $v \in Q(0; P_{\{t\}})$. By the induction hypothesis, there are xay and waz in $\tau^*(\{a\})$ with $\nu(xay) = c_T + u \in L_T$ and $\nu(waz) = c_t + v \in L_{\{t\}}$. Then $w = xwazy \in \tau^*(\{a\})$ and $\nu(\bar{w}) = \nu(xay) + \nu(waz) - 1 = a$.

Corollary 2. *Simple nested iterated substitution preserves slip languages.*

Proof. Let τ^* be a simple n.i.s. defined on Σ_1^*. For $a \in \Sigma_1$, let $\tau_1(a) = \tau^*(\{a\})$. If each $\tau(a)$ is a slip language, then each $\tau_1(a)$ is a slip language by Lemma 4. If $L \subseteq \Sigma_1^*$ is a slip language, then $\tau^*(L) = \tau_1(L)$ is a slip language by Lemma 3.

Now we are ready for our main result.

Theorem 1. *The family of slip languages is closed under nested iterated substitution.*

Proof. For a nested iterated substitution τ^*, call the number of symbols a for which $\tau(a) \neq \{a\}$ the "iteration number" of τ. We proceed by induction on the iteration number. If the iteration number of τ is one, the corollary to Lemma 4 gives the desired result.

Suppose we have the theorem for iteration number $l \geq 1$ and τ has iteration number $l+1$. Let τ be defined on Σ_1^*, and let $a \in \Sigma_1$ such that $\tau(a) \neq \{a\}$. Let $\tau_1(a) = \{a\}$, $\tau_2(a) = \tau_1^*(\tau(a))$, and $\tau_1(b) = \{b\}$, $\tau_2(b) = \{b\}$ for $b \neq a$. Then τ_1 has interation number l and τ_2 has iteration number one. If $\tau(b)$ is a slip language for all $b \in \Sigma_1$, then the same is true of $\tau_1(b)$ and $\tau_2(b)$ by the induction hypothesis.

For any $w \in \Sigma_1^*$, we claim that $\tau^*(\{w\}) = \tau_2^*(\tau_1^*(\{w\}))$. Obviously, $\tau_2^*(\tau_1^*(\{w\})) \subseteq \tau^*(\{w\})$. The reverse inclusion follows by induction on the number of substitutions for a in a word of Σ_1^*. If x is obtained from w with no substitutions of a, $x \in \tau_1^*(\{w\})$. If x is obtained from w by $k+1$ substitutions, then $x = uvz$, where uaz is obtained from w by at

most k substitutions for a and y is obtained from a with no further a-substitutions. So y is in $\tau_1^*(\tau(a)) = \tau_2(a)$. Thus if $uaz \in \tau_2^*(\tau_1^*(\{w\}))$, x is in $\tau_2(\tau_2^*(\tau_1^*(\{w\}))) = \tau_2^*(\tau_1^*(\{w\}))$. Hence $\tau^*(\{w\}) \subseteq \tau_2^*(\tau_1^*(\{w\}))$.

If L is a slip language, $\tau^*(L) = \tau_2^*(\tau_1^*(L))$ is a slip language by the induction hypothesis.

The proof of Theorem 1 yields as a corollary:

Corollary 3. *A family of languages is closed under nested iterated substitution if and only if it is closed under simple nested iterated substitution.*

Theorem 2. *If L is a nontrivial slip family of languages closed under intersection with regular sets and finite substitution, the least superAFL containing L is a slip family.*

Proof. $\hat{F}(L)$ is a slip family by Lemma 1. Every member of $\hat{S}^*(L) = \hat{S}^*(\hat{F}(L))$ can be represented as $\tau^*(L) \cap \Sigma_1^*$ for some $L \in \hat{F}(L)$, some nested iterated substitution τ^*, and some vocabulary Σ_1 (see [2]). By Theorem 1, $\tau^*(L)$ is a slip language; by Lemma 2, $\tau^*(L) \cap \Sigma_1^*$ is a slip language. Hence $\hat{S}^*(L)$ is a slip family.

Corollary 4. *The context-free languages form a slip family.*

Proof. The family of context-free languages is the least superAFL containing the regular sets (see [2]).

We conclude by exhibiting an infinite hierarchy of slip superAFL's. For each $n \geq 1$, let $a_1, \ldots, a_n$ be distinct and let $L_n = \{a_1^m \ldots a_n^m \mid m \geq 1\}$. For each n, $\hat{F}(L_n)$ is a slip family properly contained in $\hat{F}(L_{n+1})$ (see [1]).

For any family of languages L, let $\hat{S}(L)$ denote the least full AFL containing L and closed under substitution. Theorem 4.6 of [2] says that if L is in $\hat{S}^*(L)$–$\hat{S}(L)$, then L contains $\{uv^m wx^m z \mid m \geq 0\}$ for some words u, v, w, x, z with $vx \neq e$. Hence, if $n \geq 3$ and $L = \{a_1^{f(m)} a_2^m a_3^{f(m)} a_4^{f(m)} \ldots a_n^{f(m)} \mid m \geq 1\}$ for any one-one function f from N into N, then L can not belong to $\hat{S}^*(L)$–$\hat{S}(L)$. It is shown on p. 27 of

[2] for $n = 3$ that if $L \in \hat{S}^{\infty}(L)$, then L is in $\hat{F}(L)$; the proof obviously extends to $n \geq 3$.‡ Thus if L is in $\hat{S}^{\infty}(L)$, then L is in $\hat{F}(L)$.

If we take $f(m) = m$ and $L = \{L_n\}$, we see that for $n \geq 2$, $L_{n+1} \in \hat{S}^{\infty}(\{L_n\})$ if and only if $L_{n+1} \in \hat{F}(L_n)$. Hence, for $n \geq 2$, $\hat{S}^{\infty}(\{L_n\})$ is a slip superAFL properly contained in $\hat{S}^{\infty}(\{L_{n+1}\})$. Obviously $\hat{S}^{\infty}(\{L_1\}) = \hat{S}^{\infty}(\{L_2\})$ and $\hat{S}^{\infty}(\{L_2\})$ is the family of context-free languages.

References

[1] S. Ginsburg and E.H. Spanier, AFL With the semilinear property, SDC document TM-738/057/000, February, 1970.

[2] S. Greibach, Full AFL's and nested iterated substitution, Inform. Control 16 (1970) 7–35.

[3] J. Gruska, A characterization of context-free languages, J. Computer Systems Sci, to appear.

[4] I.P. McWhirter, Context-free expressions, Research Report CSRR-2016, University of Waterloo, June, 1970.

[5] R.J. Parikh, On context-free languages, J. Assoc. Computer Mach. 13 (1966) 570–581.

‡ In [2] the details are given in the case $f(n) = 2^n$; it is observed that the proof obviously extends to all one-one functions.