A STRONG PUMPING LEMMA
FOR CONTEXT-FREE LANGUAEES

David S. Wise

Computer Science Department
Indiana University
Bloomington, Indiana 47401

TECHNICAL REPORT NO. 4

A STRONG PUMPING LEMMA
FOR CONTEXT-FREE LANGUAGES

DAVID S. WISE

REVISED: APRIL, 1976

A Strong Pumping Lemma
For Context-Free Languages

David S. Wise

Computer Science Department

Indiana University

Bloomington, Indiana 47401

Abstract.

A context-free language is shown to be equivalent to a set of sentences describable by sequences of strings related by finite substitutions on finite domains, and vice-versa. As a result, a necessary and sufficient version of the classic pumping lemma is established. This result provides a guaranteed method of proving that a language is not context-free when such is the case. An example is given of a language which neither the classic pumping lemma nor Parikh's theorem can show to be non-context-free, although Ogden's lemma can. The main result also establishes $\{a^n b a^{mn}\}$ as a language which is not in the Boolean closure of deterministic context-free languages.

Introduction

One of the most useful results about regular languages is Nerode's Theorem [4], which yields a "sure-fire" scheme for proving either that a language is regular (by presenting a finite state automaton) or that it is not (by violating any finite right congruence on Σ^*).

The standard technique for establishing that a language is context-free is to present a context-free grammar which generates it or a pushdown automaton which accepts it. If it is not context-free, the classic pumping lemma [2] or Parikh's Theorem [7] often can establish the fact, but they are not guaranteed to do so, as will be seen. The characterization of context-free languages by non-deterministic pushdown automata does not solve the problem because of the difficulties in establishing constraints on arbitrary non-deterministic computations.

In this paper context-free languages are characterized by three finite substitutions on a finite domain (closely related to self-embedding non-terminals), such that a sentence is in a language precisely when a finite sequence of strings exists which are related by these substitutions in a manner reminiscent of the pumping lemma (Property 3 below). The domain is analogous to the finite set of partition blocks in Nerode's Theorem. The main theorem (Theorem 2) also establishes a form of the pumping lemma applied to sentential forms (Property 2) as equally powerful.

Two applications are presented which demonstrate the power of the results. Theorem 4 establishes

$$\{a^b b^q c^r \mid p,q,r \geq o \quad \text{and} \quad p \neq q \neq r \neq p\}$$

a non-context-free language using Property 2, whereas the classic pumping lemma and Parikh's Theorem fail to do so. Theorem 5, which is not directly obtainable from characterizations of context-free languages in terms of grammars or machines, states that $\{a^n b a^{mn}\}$ is not expressible as a finite intersection of context-free languages. It was a corollary to Theorem 5, the fact that this language is not in the Boolean closure of deterministic languages, which originally motivated this work.

Definitions.

The notation generally follows Aho and Ullman [1]. If Σ is a vocabulary, Σ^* denotes the set of strings on Σ, and Σ^+ denotes the set of non-empty strings on Σ. A grammar is a quadruple (N,Σ,P,S) where N is a set of non-terminals, Σ is a terminal alphabet such that $N \cap \Sigma = \emptyset$, $S \notin N \cup \Sigma$ is the start symbol, and $P \subset (N \cup \{S\})^+ \times (N \cup \Sigma)^*$ is the set of productions. A grammar is context-free if $P \subset (N \cup \{S\}) \times (N \cup \Sigma)^*$.

Lower case Roman letters denote characters in Σ if early in the alphabet and strings in Σ^* if at the end of the alphabet. Upper case Roman letters usually denote characters in N and upper case Greek letters will denote auxiliary alphabets. Lower case Greek letters denote arbitrary strings. Of special note is e , denoting the empty string. The length of a string α is written $|\alpha|$; $|e| = 0$.

The derives relation applies between two strings, $\alpha \underset{G}{\Rightarrow} \beta$, when a production of G applies to α and results in β , Often a production $(A,\beta) \in P$ is displayed as $A \to \beta$, with G understood. A derivation of σ_n from σ_0 is a sequence of strings $\sigma_0, \sigma_1, \ldots, \sigma_n$ such that $\sigma_{i-1} \Rightarrow \sigma_i$ for all $0 < i \leq n$. The transitive closure of $\Rightarrow$ is denoted by $\overset{+}{\Rightarrow}$, and its reflexive transitive closure is denoted by $\overset{*}{\Rightarrow}$. Note that if we say $A \overset{*}{\underset{G}{\Rightarrow}} \sigma$ then the derivation of σ can proceed without regard to the context in which A appears. If, however, $\delta_1 A\delta_2 \overset{*}{\underset{G}{\Rightarrow}} \delta_1\sigma\delta_2$ it is not necessarily true that $A \overset{*}{\underset{G}{\Rightarrow}} \sigma$. The set of sentential forms of G, denoted $SF(G)$, is $\{\sigma | S \overset{*}{\underset{G}{\Rightarrow}} \sigma\}$. The language of G, denoted $L(G)$, is $SF(G) \cap \Sigma^*$.

A finite substitution, f, is a mapping of a finite set onto finite subsets of Δ^* for some finite set Δ. The mapping f may be extended to strings in the natural manner: $f(e) = e$ and $f(A\alpha) = f(A)f(\alpha)$ for $A \in \Gamma, \alpha \in \Gamma^*$.

A set, S, of n-tuples of non-negative numbers is said to be linear if there is an integer $k \geq 0$ and n-tuples $v_0, \ldots, v_k$ such that $S = \{v_0 + \Sigma_{i=1}^{k} (m_i v_i) | m_i \geq 0$ are integers$\}$. A set of n-tuples is semi-linear if it is a finite union of linear sets. A Parikh mapping [9], q, is a mapping of $z \in \Sigma^*$ into a $|\Sigma|$-tuple of non-negative integers defined by $q(z) = (\#_{a_1}(z), \ldots, \#_{a_{|\Sigma|}}(z))$ where $\#_{a_i}(z)$ is the number of times $a_i \in \Sigma$ occurs in z. For $L \subset \Sigma^*$, define $q(L) = \{q(z) | z \in L\}$.

A non-terminal A is cyclic if $A \overset{+}{\Rightarrow} A$ and any derivation by which $A \overset{+}{\Rightarrow} A$ is a cycle. Any derivation including a cycle can be trivially shortened.

A non-terminal A is self-embedding in a context-free grammar G if $A \overset{+}{\Rightarrow} \beta A\gamma$, where $\beta\gamma \neq e$. Other authors restrict $\beta \neq e \neq \gamma$. The new definition specifies a somewhat larger class of non-terminals, elsewhere described as "recursive but not because of a cycle," which characterize grammars that are necessarily context-free, as we shall see. A production $A \rightarrow \alpha$ is said to be self-embedded at the jth step of a derivation $\sigma_1 \Rightarrow \sigma_2 \Rightarrow \cdots \Rightarrow \sigma_n$ if for $1 \le i < j < n$: $\sigma_i = \delta_1 A \delta_2$, $\sigma_j = \delta_1 \beta A \gamma \delta_2$, and $\sigma_{j+1} = \delta_1 \beta \alpha \gamma \delta_2$ where the productions applied in the ith through the (j-1)st step effect the self-embedding $A \overset{+}{\Rightarrow} \beta A \gamma$, $\beta\gamma \ne e$. Intuitively, a production is self-embedded if its left part has already generated a self-embedding at that point in the sentential form. A self-embedding chain from A is a derivation $A \overset{+}{\Rightarrow} \beta A \gamma$ with no self-embedded productions. The derivation tree of any self-embedding chain is bounded in depth by $|N|$ and in degree of any node by the length of the longest production, so for any context-free grammar there are only a finite number of them.

Results

The "reflex" tactic for proving that a language is not context-free is to obtain a contradiction of Bar-Hillel's "pumping" lemma [2] (the "classic pumping lemma"), Parikh's "semilinear" theorem [7], or Ogden's lemma [6] (an extended version of the classic pumping lemma). Often the given language is intersected with a regular set or transformed by a gsm mapping before one of these techniques is applied. If a language is context-free then the conditions stated in the pumping lemma, Parikh's theorem, and Ogden's lemma are necessarily satisfied, but none of them are known to guarantee that a language is context-free. Therefore, there is no guarantee that these statements will generate a contradiction if it is improperly assumed that a language or its transformed image is context-free. Ogden's Lemma is, however, more powerful than the classic pumping lemma or Parikh's theorem and may characterize context-free languages.

Theorem 1 (Ogden's lemma) [6]. For each context-free grammar $G = (N,\Sigma,P,S)$ there is an integer k such that for any word $z \in L(G)$, if any k or more distinct positions in z are designated as distinguished, then there is some $A \in N$, and strings $u,v,w,x,y \in \Sigma^*$ such that

(i) $S \overset{*}{\Rightarrow} uAy$; $A \overset{+}{\Rightarrow} vAx$; $A \overset{+}{\Rightarrow} w$; $uvwxy = z$.

(ii) $w \neq e$ contains at least one distinguished position.

(iii) Either u and v both contain distinguished positions or x and y both do.

(iv) vwx contains at most k distinguished positions.

I know of no non-context-free language which displays the property cited for any of its grammars, but it is not known whether satisfying (i) - (iv) of Theorem 1 is sufficient to establish that a language is context-free. The emphasis of the theorem is on "distinguished positions", yet it is unclear why a grammar which satisfies (i) - (iv) might necessarily describe a context-free language.

In an attempt to capture the essence of the context-free language property, we shall prove the following three statements to be equivalent.

1. L is context-free.
2. There is an unrestricted grammar G and an integer k such that $L = L(G)$ and when $\sigma \in SF(G)$, $|\sigma| > k$, then σ may be rewritten as $\sigma = \upsilon\nu\omega\chi\psi$ where $\omega \neq e$, $\upsilon \neq e$ or $\chi \neq e$, $|\nu\omega\chi| \leq k$, and there is a non-terminal A in G such that $S \overset{*}{\Rightarrow} \nu A\psi$, $A \underset{G}{\overset{+}{\Rightarrow}} \nu A\chi$, and $A \underset{G}{\overset{*}{\Rightarrow}} \omega$.
3. $L \in \Sigma^*$ and there exist

a finite alphabet Γ and a distinguished $S \notin \Gamma$, disjoint from Σ ,

a substitution, h , mapping $\Gamma \cup \{S\}$ onto finite subsets of $(\Gamma\cup\Sigma)^*$ whose domain is extended to Σ by defining $h(a) = \{a\}$ for $a \in \Sigma$ and extended thence to strings on $\Sigma \cup \Gamma \cup \{S\}$ in the usual manner,

and two substitutions, f and g , mapping $\Gamma \cup \{S\}$ onto finite subsets of $(\Gamma\cup\Sigma)^*$ such that $e \notin f(C)g(C)$ for all $C \in \Gamma$ but $f(S) = \{e\} = g(S)$.

such that whenever $z \in L$ there is a finite sequence $\sigma_0, \sigma_1, \sigma_2, \ldots, \sigma_m$ of strings in $(\Sigma\cup\Gamma)^*$ such that $S = \sigma_0$, $z = \sigma_m$, and for $1 \leq j \leq m$, σ_j may be rewritten $\sigma_j = \upsilon\nu\omega\chi\psi$, $\sigma_{j-1} = \upsilon C\psi$ for some $C \in \Gamma \cup \{S\}$, where $\nu \in f(C)$, $\omega \in h(C)$, $\chi \in g(C)$, and $h(\upsilon\nu^i C\chi^i\psi) \cap \Sigma^* \subset L$ for all $i \geq 0$.

Property 3 is somewhat unwieldy but does avoid the terminology of grammars and derivations. The flavor of a context-free grammar shows through; self-embeddings are the essence of other characterizations as well. We may consider $C \in \Gamma$ to be a triple (β, A, γ) which represents a self-embedding chain $A \overset{+}{\underset{G}{\Rightarrow}} \beta A\gamma$ in some context-free grammar G for L . If p is a substitition mapping each nonterminal A into the set of triples which represent self-embedding chains for A in G then $f(C) = p(\beta)$, $g(C) = p(\gamma)$, and $h(C) = \{p(\sigma) | A \overset{*}{\underset{G}{\Rightarrow}} \sigma$ with no self-embedded productions$\}$. Property 3 concentrates our attention upon the finiteness of self-embedding chains, which is somewhat analogous to the finiteness of congruence classes in Nerode's theorem [8].

<u>Theorem 2 (Strong pumping lemma)</u>. Properties 1, 2, and 3 above are equivalent.

<u>Proof</u>. 1 ⇒ 2 . If $G = (N, \Sigma, P, S)$ is context-free, construct

N' , Σ' , and P' by priming all characters in the vocabulary. Then $G' = (N' \cup \Sigma', N \cup \Sigma, P' \cup \{A' \to A | A \in N \cup \Sigma\}, S')$ is context-free and $L(G') = SF(G)$. If $\sigma \in SF(G)$ then Theorem 1 can be applied (using G') whenever $|\sigma| \geq k$ if all positions are distinguished. The trivial homomorphism from $SF(G')$ to $SF(G)$ establishes $\underline{\underline{2}}$.

$\underline{\underline{2}} \Rightarrow \underline{\underline{3}}$. Given k and $G = (N,\Sigma,P,S')$ as described in $\underline{\underline{2}}$, for every $A \in N$ define $p(A) = \{(\beta,A,\gamma) | \beta\gamma \in (N \cup \Sigma)^+ , A \overset{+}{\underset{G}{\Rightarrow}} \beta A \gamma$, and $0 < |\beta\gamma| \leq k\}$. The set $p(A)$ includes all self-embedding chains on A which are necessary to enforce Property $\underline{\underline{2}}$. Some other self-embedding chains from G not necessary to Property $\underline{\underline{2}}$ (perhaps because G is ambiguous) may be excluded by the length restriction of k . It is also possible that G allows derivation steps which are not reflected in Property $\underline{\underline{2}}$ and therefore do not contribute to $p(A)$ for any non-terminal A . Since Property $\underline{\underline{2}}$ applies to every sentential form, however, we shall be able to describe some derivation for every sentence in L in terms of the p mappings. Since G is unrestricted, $p(A)$ may not be effectively constructable, but it does exist and is finite because of the bound k .

Define $S = (e,S',e)$ and $\Gamma = \bigsqcup_{A \in N} p(A)$. Γ is clearly a finite set. Define $p(a) = a$ for $a \in \Sigma$ and extend p to a length-preserving string substitution on $(N \in \Sigma)^*$ in the natural manner. The substitutions f , g , and h defined as follows are also finite:

$$h((\beta,A,\gamma)) = \bigsqcup_{\substack{|\omega| \leq k \\ A \overset{*}{\underset{G}{\Rightarrow}} \omega}} p(\omega)$$

$$h(a) = \{a\} \text{ for } a \in \Sigma ,$$

$$f((\beta,A,\gamma)) = p(\beta) \text{ , and}$$

$$g((\beta,A,\gamma)) = p(\gamma) \text{ .}$$

Let $z \varepsilon L$. If $|z| \le k$ we have $z \varepsilon h(S)$ and Property $\underline{\underline{3}}$ is satisfied with $m = 1$.

Suppose that $|z| > k$. Beginning with z apply Property $\underline{\underline{2}}$ repeatedly to get a sequence of sentential forms $z = \zeta_m, \cdots, \zeta_1$ such that $|\zeta_1| \le k$, $S' \overset{+}{\underset{G}{\Rightarrow}} \zeta_1 \overset{+}{\underset{G}{\Rightarrow}} \cdots \overset{+}{\underset{G}{\Rightarrow}} \zeta_m = z$ and $\zeta_j = \upsilon_j \nu_j \omega_j \chi_j \psi_j$ where, for some $A_j \in N$, $\zeta_{j-1} = \upsilon_j A_j \psi_j$, $A_j \overset{+}{\underset{G}{\Rightarrow}} \nu_j A_j \chi_j$, and $A_j \overset{*}{\underset{G}{\Rightarrow}} \omega_j$ for all $1 < j \le m$. Moreover, Property $\underline{\underline{2}}$ assures the existence of such a sequence with $|\nu_j \chi_j| > 0 < |\omega_j|$ and $|\nu_j \omega_j \chi_j| \le k$, so $|\zeta_{j-1}| < |\zeta_j|$ and the sequence is finite: $m < |z|$. If we assume that there exists a string $\sigma_j =$ $\bar{\upsilon}_j \bar{\nu}_j \bar{\omega}_j \bar{\chi}_j \bar{\Psi} \in p(\zeta_j)$ such that $\bar{\upsilon}_j \in p(\upsilon_j)$; $\bar{\nu}_j \in p(\nu_j)$; $\bar{\omega}_j \in p(\omega_j)$; $\bar{\chi}_j \in p(\chi_j)$; and $\bar{\Psi}_j \in p(\psi_j)$ (and this is trivially true for $j = m$) then we may easily construct $\sigma_{j-1} = \bar{\upsilon}_j\ (\nu_j, A_j, \chi_j)\ \bar{\Psi}_j \in p(\zeta_{j-1})$. We still have $\bar{\upsilon}_j \in p(\upsilon_j)$ and $\psi_j \in p(\psi_j)$, and by definition $(\nu_j, A_j, \chi_j) \in p(A_j)$ because $|\nu_j \omega_j \chi_j| \le k$ implies $|\nu_j \chi_j| < k$.

Moreover, $\bar{\nu}_j \in f((\nu_j, A, \chi_j)) = p(\nu_j)$

$\bar{\chi}_j \in g((\nu_j, A, \chi_j)) = p(\chi_j)$

and $\bar{\omega}_j \in h((\nu_j, A_j, \psi_j))$.

This last fact holds since $A_j \overset{*}{\Rightarrow} \omega_j$, $|\omega_j| < k$, and $\bar{\omega}_j \in p(\omega_j)$.
At each step we may identify C as the triple (ν, A_j, χ_j) and ω as the remaining substring of σ_j to see that the rewriting $\sigma_j = \overline{\upsilon\nu\omega\chi\psi}$ and $\sigma_{j-1} = \bar{\upsilon} C \bar{\psi}$ required by Property $\underline{\underline{3}}$ is indeed possible. Finally set $\sigma_0 = S$ and note that σ_1 is necessarily in $h(S)$ because $|\sigma_1| = |\zeta_1| \le k$ and that $f(S) = e = g(S)$.

Furthermore, every string z which can be obtained from S via a finite number of f,g,h substitutions is in L . If $|z| > k$ and a sequence $S = \sigma_0, \cdots, \sigma_m = z$ met the constraints of Property <u>3</u> with f , g and h as defined above, then it is easy to see that Property <u>2</u> is also met m times. By establishing that the trivial inverse p image of each σ_j is a sentential form of the original grammar G , we show that z is necessarily in L .

$3 \Rightarrow 1$. Suppose that Σ, Γ, S, f, g and h are given as in 3, defining a language L. Let $G = (\Sigma,\Gamma,P,S)$ where P is constructed as follows:

$$P = \bigcup_{A \in \Gamma \cup \{S\}} \{A \to \beta\alpha\gamma \ , \ A \to \beta A\gamma \mid \beta \in f(A) \ , \ \alpha \in h(A) \ , \ \gamma \in g(A)\}$$

Now suppose $z \in L(G)$ and consider its derivation $S \underset{G}{\Rightarrow} \sigma_1 \Rightarrow \cdots \Rightarrow \sigma_m = z$. This sequence of sentential forms satisfies the requirements of Property 3 on the sequence of σ_i, so that $z \in L$. On the other hand, if $z \in L$ the sequence of σ_i (which necessarily exists) describes a derivation of z in G. Hence $L(G) = L$, so L is context free. ■

Note that the grammar constructed immediately above may have one e-production, because there is no restriction preventing $e \in h(C)$ for $C \in \Gamma$. In particular, when $e \in L$ then $e \in f(S)h(S)g(S)$.

Corollary 1. (Pumping lemma) [2]. If L is context-free, then there exist integers m and n such that when $z \in L$, $|z| > m$ then z may be written $z = uvwxy$ where $|vwx| \leq n$, $vx \neq e$, and $uv^i wx^i y \in L$ for all $i \geq 0$.

Proof. Just as Ogden [6] proved this from Theorem 1, apply Property 2 with $k = m = n$. ■

Theorem 3 [7]. If L is context-free then the Parikh mapping of L, $q(L)$, is semi-linear.

This result has been elegantly proved in a more general form [5], but it is worthwhile noting that the classical proof (e.g., [9]) hinges precisely on Γ described in Property 3. That proof can be abbreviated by a modification of the previous construction.

Applications.

Theorem 2 guarantees us a scheme for proving a language is not context-free. The first example illustrates that power, using Property 2 on a case for which Corollary 1 and Theorem 3 are useless.

The language $L_1 = \{a^p b^q c^r | p \neq q \neq r \neq p\}$, suggested by a referee, is not a context-free language, but it is impossible to establish that fact using these techniques although Theorem 1 does apply. It is important to realize that gsm mappings and intersections with regular sets do not usefully transform L_1 . Its structure is so simple that these transformations yield trivially context-free languages, or languages even more complex than L_1 .

Theorem 4. L_1 is not context-free, but its Parikh mapping is semi-linear, and for all $z \in L_1$, $|z| > 3$ may be rewritten as $z = uvwxy$ where $|vwx| \leq 3$, $vx \neq e$ and $uv^i wx^i y \in L_1$ for all $i \geq 0$.

Proof. The Parikh mapping of L_1 is a union of six linear sets of triples, corresponding to the six ways of ordering three distinct integers. The linear set corresponding to the case in which the integers (p,q,r) are in decreasing order is generated by $\{(2,1,0) + i(1,1,1) + j(1,1,0) + k(1,0,0) | i,j,k \geq 0\}$. The other five linear sets are generated by uniformly permuting the co-ordinates of all vectors in this set. Each $a^p b^q c^r \in L_1$ is a member of the linear set identified by the sorting of (p,q,r) .

The classic pumping criterion always applies to $a^{j[a]}b^{j[b]}c^{j[c]} \in L_1$ when $j[a] + j[b] + j[c] > 3$. Choose $t \in \Sigma = \{a,b,c\}$ such that $j[t]$ is largest, and then choose $1 \leq k \leq 3$ so that $k \neq j[t]-j[s]$ for all $s \in \Sigma$. Since $|z| > 3$ it follows that $k \leq j[t]$. Let $v = t^k$, $w = e$, $x = e$, and u and y be appropriate so that $uvy = z$. It follows easily that $|vwx| \leq 3$ and $uv^iwx^iy \in L_1$ for all $i \geq 0$.

Finally we must establish that L_1 is not context-free. Theorem 1 yields an easy contradiction to the assumption that it is by considering $a^k b^{k+k!} c^{k+2k!}$ with the first k positions distinguished. Any factorization must pump a's, or a's and b's, or a's and c's. If there are precisely q a's in vx then $1 \leq q \leq k-2$ and q divides $k!$ If $b \in vx$ then let $i = 1 + (2k!/q)$; otherwise let $i = 1 + (k!/q)$. In both cases $S \overset{*}{\Rightarrow} uv^iwx^iy \notin L_1$; yielding a contradiction. ■

Ogden's Lemma uses "distinguished positions" to isolate pumping to a particular part of the sentence, avoiding effects other pumpings which may be possible due to self-embedding chains in remote parts of the derivation tree. Possible pumping affecting only one part of the sentence (the suffix $b^{k+k!}c^{k+2k!}$ in the above proof) can be ignored while the desired pumping can be studied by distinguishing characters somewhere else (in the prefix a^k above). A proof that L_1 is not context-free using Theorem 2 requires consideration of the effects of two pumpings, which we shall select from three which are certainly possible in deriving a sentence of length greater than $3k$.

<u>Proof</u> that L_1 is not context-free using Property $\underline{\underline{2}}$. Let G be an unrestricted grammar (with $\Sigma = \{a,b,c\}$) possessing Property $\underline{\underline{2}}$ for the constant k . Suppose Property $\underline{\underline{2}}$ were applied (in parallel) to all sentences z in L_1 of length longer than k , and to all sufficiently long sentential forms uncovered as a consequence of applying it. In that way all factorizations $\upsilon\nu\omega\chi\psi$ of sufficiently long but useful sentential forms in L_1 could be identified. We are interested in all of the possible candidates for $\nu\omega\chi$ in these factorizations, which form a subset of $(N\cup\Sigma\cup\{e\})^k$ since $|\nu\omega\chi| \leq k$. We are particularly interested in those factorizations with ν or χ in a^+ , b^+ or c^+ , noting that if ν or χ is in Σ^+ then it is necessarily in one of these three languages.

We can bound the number of terminal strings v , w , and x derivable without subsequent self-embeddings from ν , ω , and

χ (respectively) as implied by applications of Property $\underline{2}$ in identifying any factorization: $\bar{u}\nu\omega\chi\bar{\psi}$. Such derivations from ν , ω , or χ can only include rewritings of the form $\bar{A} \overset{*}{\Rightarrow} \bar{\omega}$ (for some $\bar{A}$, $\bar{\omega}$), since ones of the form $\bar{A} \overset{+}{\Rightarrow} \bar{\nu}\bar{A}\bar{\chi}$ are excluded and no other sort are implied by applying Property $\underline{2}$. Let n be a common multiple of the lengths of all candidates for ν,χ and $\nu\chi$ and consider $z = a^k b^{k+n} c^{k+2n} \in L_1 \subset SF(G)$.

The pumping of Property $\underline{2}$ applies at least <u>thrice</u> in some derivation of z because of its length, and we shall choose two pumpings upon which to base a contradiction to the assumption that L_1 is context-free. Let us require that each $\nu\chi$ for the three applications be such that one contains an a, one a b, and one a c . (Since there are at least k of each this can be forced.) Let us associate A_s , ν_s , ω_s , χ_s for $s \in \Sigma$ with the application which satisfies this constraint: $s \in \nu_s \chi_s$. It is possible that the labelling is not unique: e.g. $A_a = A_b$, $\nu_a = \nu_b$, $\omega_a = \omega_b$, $\chi_a = \chi_b$ is possible.

Let (r,s,t) stand for any permutation of the triple (a,b,c) in the following argument. It is impossible that both A_s and A_t are derivable from $\nu_r\chi_r$ in the application pattern we identified for z. If this were possible, pumping of $\nu_r{}^i\omega_r\chi_r{}^i$ would either introduce multiple occurrences of A_s or A_t with r's derivable between (e.g. ν_r or $\chi_r \overset{+}{\Rightarrow} \ldots r \ldots A_s \ldots$; ν_r or $\chi_r \overset{+}{\Rightarrow} \ldots A_s \ldots r \ldots$) or in case ν_r or $\chi_r \overset{+}{\Rightarrow} A_s$ where either $A_s \overset{+}{\Rightarrow} \ldots s \ldots A_t \ldots$ or $A_t \overset{+}{\Rightarrow} \ldots A_t \ldots s \ldots$ then that pumping would introduce a sequence of s's (or t's) with A_t (respectively A_s) interspersed. In all these events, since

$A_d \to \ldots d \ldots$ for $d \in \Sigma$, we can derive a sentence not in $a^*b^*c^* \supset L_1$ by pumping A_r and reapply Property 2 derivations. The argument holds regardless of permutation.

As a result, at least two of our three applications of Property 2 are such that $\nu_s \chi_s$ does not derive A_t and $\nu_t \chi_t$ does not derive A_s for $s \neq t$ both in Σ. We can even force the following to be true by choosing three appropriate instances of Property 2 which arise early in analyzing z :

$$[(\nu_s \in s^+ \text{ and not } \chi_s \stackrel{*}{\Rightarrow} \epsilon\ s^+) \text{ or } (\text{not } \nu_s \Rightarrow \epsilon\ s^+ \text{ and } \chi_s \in s^+) \text{ or } \nu_s \chi_s \in s^+]$$

and similarly for t instead of s . Let p be the number of s's in $\nu_s \chi_s$ and q be the number of t's in $\nu_t \chi_t$. By the definition of n both p and q divide n .

Now if $s = a$, pump $A_s \Rightarrow \nu_s \omega_s \chi_s$ a total of $2n/p$ times: $A_a \stackrel{+}{\Rightarrow} \nu_a^{(2n/p+1)} \omega_a \chi_a^{(2n/p+1)}$ which has $2n$ more a's than $\nu_a \omega_a \nu_a$. If $s = b$ pump n/p times adding n b's; if $s = c$ do not pump adding no new occurrences of c . Similarly if $t = a$ pump $A_t \stackrel{+}{\Rightarrow} \nu_t A_t \chi_t$ $2n/q$ times; if $t = b$ pump n/q times; if $t = c$ do not pump $A_c \stackrel{+}{\Rightarrow} \nu_c A_c \chi_c$ at all. Since A_s and A_t appear independently of each other in the derivation we have constructed, neither pumping creates new occurrences of the other non-terminal and so the only effect is to derive a new terminal string in the language of G .

In any event no new occurrrences of c are added to G , but either $2n$ a's or n b's are added. Then we have either $a^{k+2n} b^{k+n} c^{k+2n} \in L_1$ or $a^k b^{k+2n} c^{k+2n} \in L_1$ which are both contradictions. So L_1 must not be context-free. ■

The esoteric flavor of Property 3 is particularly useful for theorems like the following.

<u>Theorem 5</u>. $L_2 = \{a^n ba^{mn} | m,n > 0\}$ cannot be expressed as the intersection of any finite number of context-free languages.

<u>Proof</u>. Suppose L_2 were expressible by a finite number of conjuncts, each of which is context-free and a subset of the regular language a^*ba^* . Characterize each of these hypothetical conjuncts by Property 3, let f , g , and h be the union of all the corresponding finite substitutions, and let Γ be the union of their domains. Define r to be a common multiple of the set of integers $\{|f(c)g(c)|$ for $C \in \Gamma\}$. Let p and q be two prime numbers : $p,q > r$. Now let $z = a^p ba^{pq} \in L_2$ so that z is in each conjunct language.

Apply Property 3 to z with respect to an arbitrary conjunct language. It is necessary that the sequence of $S = \sigma_0, \ldots, \sigma_m = z$ described in Property 3 have a largest n such that $\sigma_n = \upsilon\nu\omega\chi\psi$, $\sigma_{n-1} = \upsilon C \psi$, $\nu \in f(C)$, $\chi \in g(C)$, and either $b \in h(\upsilon)$ with $\nu\chi \in a^+$ or $b \in h(\omega)$ with $\nu = e$, $\chi \in a^+$. The "pumping" must sometime apply to the right of (the preimage under h of) b because pq is so large. Applying the rewritings as on $\sigma_{n+1}, \ldots, \sigma_m$ we see that each conjunct language has a subset of the form

$$\{a^p b a^{pq+(i-1)t} | i \geq 0\}$$

where t is a positive integer reflecting the length of $\nu\chi \in a^+$. (The value of $t \neq 0$ will be $|f(C)g(C)|$ for that C.) Although t varies with different languages, t must divide r. For varying choice of i we can force $a^p b a^{pq+r} \in L_2$. However p cannot evenly divide $pq + r$, so $a^p b a^{pq+r} \notin L_2$, yielding the desired contradiction. ∎

Corollary 2. L_2 is not a member of the Boolean closure of the deterministic context-free languages [4].

Proof. If L_2 were in that class then it could be expressed as some Boolean combination of deterministic languages in conjunctive normal form. Each conjunct would necessarily be context-free because the class of deterministic languages is closed on complementation, because each deterministic language is context-free, and because context-free languages are closed on union. Theorem 5 does the rest. ■

Conclusions.

The examples of the last section demonstrate that Corollary 1 and Theorem 3 do not characterize context-free languages. However, this weakness does not appear in Theorem 1 (which may indeed be necessary and sufficient). Theorem 2 shows that the essence of context-free languages is a pumping property of a finite nature which may appear at different points in the sentence. The pumping may be characterized by the self-embedding chains of a grammar. Alternatively, it may be expressed as a set of finite substitutions on a finite domain, avoiding the terminology of grammatical derivations. While Theorem 1 has this flavor, it is clouded by the power of selecting distinguished characters.

Therefore, the universal strategy for proving that a language is not context-free (when such is the case) is to assume it is characterized by Property 2 or Property 3 and search for the guaranteed contradiction.

Acknowledgment. The author is grateful for the comments of Jan van Leeuwen, of Michael Harrison, and of the referees, particularly ____________ ____________ .

References.

1. Aho, A.V., and Ullman, J.D. The Theory of Parsing Translation and Compiling 1, Parsing, Prentice-Hall, Englewood Cliffs (1972).

2. Bar-Hillel, Y.; Perles, M.; and Shamir, E. On formal properties of simple phrase structure grammars. Z. Phonetik. Sprachwiss. Kommanikai. 14 (1961), 143-172. Also in Bar-Hillel, Y. Language and Information, Addison-Wesley, Reading (1964).

3. Chomsky, N. On certain formal properties of grammars. Information and Control 2, 2 (June, 1959), 137-167.

4. Ginsburg, S., and Greibach, S. Deterministic context-free languages. Information and Control 9, 6 (December, 1966), 620-648.

5. Van Leeuwen, J. A generalization of Parikh's Theorem in formal language theory. Proc. 2nd Colloquium on Automata, Languages and Programming, Springer Lecture Notes in Computer Science 14 (1974), 17-26.

6. Ogden, W. A helpful result for proving inherent ambiguity. Math. Systems Theory 2, 3 (September, 1968), 191-194.

7. Parikh, R.J. On context-free languages. J. Assoc. Comput. Mach. 13, 4 (October, 1966), 570-581.

8. Rabin, M.O., and Scott, D. Finite automata and their decision problems. IBM J. Res. Develop. 3, 2 (April, 1959), 114-125. Also in Moore, E.F. (Ed.) Sequential Machines, Addison-Wesley, Reading (1964).

9. Salomaa, A. Formal Languages, Academic Press, New York (1973).