Printing Requisition/Graphic Services

Dept. No. 97137

Title or Description: Pumping Lemmas for Term Languages

Date: 10/3/78

Date Required: A.S.A.P. PLEASE & THANK YOU

Account: 126-6294-41

Signature: [illegible]

Signing Authority: [illegible] for [illegible]

Department: Computer Science

Room: 5180

Phone: 3143

Delivery: ☐ Mail ☐ Pick-up ☐ Via Stores ☐ Other

1. Please complete unshaded areas on form as applicable. (4 part no carbon required).
2. Distribute copies as follows: White, Canary and Pink—Printing, Arts Library or applicable Copy Centre Goldenrod—**Retain.**
3. On completion of order, pink copy will be returned with printed material. Canary copy will be costed and returned to requisitioner, **Retain as a record of your charges.**
4. Please direct enquiries, quoting requisition number, to Printing/Graphic Services, Extension 3451.

Reproduction Requirements: ☑ Offset ☐ Signs/Repro's ☐ Xerox

Number of Pages: 22

Number of Copies: 50

Type of Paper Stock: ☑ Bond ☐ Book ☐ Cover ☐ Bristol ☐ Supplied

Paper Size: ☑ 8½ x 11 ☐ 8½ x 14 ☐ 11 x 17

Paper Colour: ☑ White ☐ Other

Ink: ☐ Black

Printing: ☑ 1 Side ☐ 2 Sides

Numbering to

Binding/Finishing Operations: ☑ Collating ☑ Corner Stitching ☐ 3 Ring ☐ Tape ☐ Plastic Ring ☐ Perforating

Folding Finished Size

Cutting Finished Size

Special Instructions: 50 BACKS & 50 COVERS ENCLOSED

Cost: Time/Materials	Fun.	Prod.Un.	Prod. Cl.	Opr. No.	Mins.	Total
Signs/Repro's	1					
Camera	2					
Correcting & Masking Negatives	3					
Platemaking	4					
Printing	5					
Bindery	6			21	20	
				22	15	
Sub. Total Time						
Sub. Total Materials						
Prov. Tax						
Total						

Film Qty Size

Plates Qty Size & Type

Paper Qty Size

Plastic Rings Qty Size

Outside Services

Graphic Services May 68 482-2

PUMPING LEMMAS FOR TERM LANGUAGES*

by

T.S.E. Maibaum

Research Report CS-78-18

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

Revised

March 1978

* This research was partially supported by a University of Waterloo Research Grant and by a grant from the National Research Council of Canada.

Pumping Lemmas

T.S.E. Maibaum
Department of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada

Pumping lemmas are stated and proved for the classes of regular and context-free sets of terms. The lemmas are then applied to solve decision problems concerning these classes of sets.

0. Pumping lemmas have been produced in various versions for a number of classes of languages (Bar-Hillel, Perles and Shamir; Moore; Hayashi; Ogden). Their use is two-fold. On the one hand, they lead to algorithms for deciding certain problems about languages such as emptiness and finiteness. On the other hand, they provide an effective means of proving that some language does not belong to a certain class.

In this paper, we provide pumping lemmas for regular and context-free term grammars (Thatcher and Wright; Brainerd; Rounds; Maibaum). (The pumping lemma for regular sets is really implicit in Thatcher and Wright.) As a consequence, we can derive the effective methods outlined above. Algorithms do exist for deciding the emptiness/finiteness of context free sets of terms, but these are indirect (Rounds). They depend on algorithms to solve the same problems for indexed languages (Aho). (Note added in revision: It has been brought to my attention by A. Salomaa through K. Culik that Aho's proof of the decidability of the emptiness problem for indexed grammars is incorrect and no correct proof is known. The pumping lemma for context free term grammars can now be used to provide a proof of this important theorem.)

We begin in section 1 by introducing some algebraic concepts which we will need. We also define and state some properties of regular and context free term grammars. In section 2, the pumping lemmas are stated and proved. In section 3, these lemmas are applied in proofs of non-membership of some sets in some classes of languages.

1. We begin by introducing some essential algebraic concepts. Let $\underline{N}$ be the set of natural numbers. A ranked alphabet is a family of sets indexed by $\underline{N}$. We use the notation $\Sigma = \{\Sigma_n\}_{n \in \underline{N}}$ for ranked alphabets. If

$f \in \Sigma_n$, f is said to be of rank n. Σ is said to be finite if the (disjoint) union of $\{\Sigma_n\}_{n\in\underline{N}}$ is finite.

A Σ-algebra is a pair consisting of a set A, called the <u>carrier</u> of the algebra, and an indexed family of assignments $\alpha = \{\alpha_n\}_{n\in\underline{N}}$ such that $\alpha_n:\Sigma_n \to (A^n \to A)$. $(A^n \to A)$ is the set of n-ary functions from A^n to A. Thus, for $f \in \Sigma_n$, $\alpha_n(f) = f_A$ is a function from A^n to A. We denote the Σ-algebra with carrier A by A_Σ.

Let X be any set such that $X\cap(\cup_n\{\Sigma_n|n\in\underline{N}\}) = \phi$ and consider the set $W_\Sigma(X)$ defined by:

(0) $X \subseteq W_\Sigma(X)$;

(i) If $f \in \Sigma_n$ and $t_i \in W_\Sigma(X)$ for $1 \le i \le n$, then $ft_1 \ldots t_n \in W_\Sigma(X)$.

$W_\Sigma(X)$ is called the set of <u>expressions</u> or <u>terms generated</u> by X.

We can make the set $W_\Sigma(X)$ into the carrier of a Σ-algebra (also denoted by $W_\Sigma(X)$) by assigning to $f \in \Sigma_n$ the operation $f_{W_\Sigma(X)}(t_1,\ldots,t_n) = ft_1\ldots t_n$. Let W_Σ denote $W_\Sigma(\phi)$.

A <u>homomorphism</u> is a structure preserving mapping $\psi:A_\Sigma \to B_\Sigma$ between two Σ-algebras, i.e. $\psi(f_A(a_1,\ldots,a_n)) = f_B(\psi(a_1),\ldots,\psi(a_n))$ for all $a_1,\ldots,a_n \in A$ and $f \in \Sigma_n$.

<u>Unique Extension Lemma</u>: Given a Σ-algebra A_Σ and an assignment $\psi:X \to A$, there is exactly one extension of ψ to a homomorphism $\bar{\psi}:W_\Sigma(X) \to A_\Sigma$. In particular, there is a unique homomorphism $h_A:W_\Sigma \to A_\Sigma$. □

We now define the binary operation of <u>substitution</u> (denoted by $\circ$) on the sets $W_\Sigma(X_n)$ and $(W_\Sigma(X_m))^n$ where $X_k = \{x_1,\ldots,x_k\}$. (See also Thatcher (1970), (1972) and ADJ).

Let $t \in W_\Sigma(X_n)$, $t_j \in W_\Sigma(X_m)$ for $1 \le j \le n$. Then $\circ: W_\Sigma(X_n) \times (W_\Sigma(X_m))^n \to W_\Sigma(X_m)$ is defined by $\circ(t, <t_1,\ldots,t_n>) = \overline{[t_1,\ldots,t_n]}(t)$ where $\overline{[t_1,\ldots,t_n]}$ is the unique homomorphism obtained from the assignment $[t_1,\ldots,t_n]: X_n \to W_\Sigma(X_m)$ which assigns t_j to x_j for $1 \le j \le n$. (See the Unique Extension Lemma.) Informally, $\circ(t,<t_1,\ldots,t_n>)$ means: Substitute t_i for each occurrence of x_i in t, $1 \le i \le n$. From now on, we will adopt the infix notation $t \circ <t_1,\ldots,t_n>$ rather than the prefix notation above. We can extend substitution to a binary operation $\circ: W_\Sigma(X_n))^p \times (W_\Sigma(X_m))^n \to (W_\Sigma(X_m))^p$ with the definition $<t_1,\ldots,t_p> \circ <t_1',\ldots,t_n'> = <t_1 \circ <t_1',\ldots,t_n'>,\ldots,t_p \circ <t_1',\ldots,t_n'>>$.

Let B(A) be the set of all subsets of a set A. Given a Σ-algebra A, B(A) can easily be made into a Σ-algebra. For given $f \in \Sigma_n$ and for $1 \le i \le n$, $S_i \subseteq A$ (i.e. $S_i \in B(A)$), define $f_{B(A)}(S_1,\ldots,S_n) = \{f_A(s_1,\ldots,s_n) | s_i \in S_i$ for $1 \le i \le n\}$. Let $t \in W_\Sigma(X_n)$ and $\alpha = <\alpha_1,\ldots,\alpha_n> \in (B(W_\Sigma(X_k)))^n$. Define $t \circ \alpha = \bar{\alpha}(t)$ where $\bar{\alpha}: W_\Sigma(X_n) \to B(W_\Sigma(X_k))$ is the unique homomorphism extending the assignment which assigns to x_i the set α_i. This is sometimes called <u>non-uniform</u> substitution because different occurrences of x_i in t can be assigned different values from the i'th component of α. Suppose $\beta \in B(W_\Sigma(X_n))$. Then $\beta \circ \alpha = \{t \circ \alpha | t \in \beta\}$. Substitution is then easily extended to the case where β is a tuple of sets. If $\alpha = <\{t_1\},\ldots,\{t_n\}>$ is a tuple of singletons, then we also write $\alpha = <t_1,\ldots,t_k>$.

<u>Theorem</u> Substitution is an associative operation. i.e. let $\alpha \in (B(W_\Sigma(X_n)))^p$, $\beta \in (B(W_\Sigma(X_m)))^n$, $\gamma \in (B(W_\Sigma(X_q)))^m$. Then $\alpha \circ (\beta \circ \gamma) = (\alpha \circ \beta) \circ \gamma$.

<u>Proof</u>: This theorem is a simple consequence of the Unique Extension Lemma and a full proof can be found in ADJ. □

Let us write $\alpha(x_1,\ldots,x_n)$ to signify $\alpha \in (B(W_\Sigma(X_n)))^p$ for some p (where p is determined by the context). For example, in the context $\beta(x_1,\ldots,x_k) \circ \alpha(x_1,\ldots,x_n)$ we see p=k by the definition of substitution. Let us write $\alpha(x_1,\ldots,x_i^!,\ldots,x_n)$ to denote $\alpha \in B(W_\Sigma(X_n))$ with exactly one occurrence of x_i in any $t \in \alpha$. Finally, given $\alpha \in (B(W_\Sigma(X_n)))^k$, $\beta \in (B(W_\Sigma(X_k)))^p$, let us write $\beta(\alpha(x_1,\ldots,x_n))$ for $\beta\circ\alpha$.

A <u>context-free term grammar</u> (Rounds; Maibaum) G over the alphabet Σ is a 4-tuple (N,Σ,P,S) where:

(i) N is a finite ranked alphabet called the set of <u>non-terminals</u> of G;

(ii) Σ is a finite ranked alphabet called the set of <u>terminals</u> of G. Let $V = \{V_n\}_{n\in\underline{N}} = \{N_n \cup \Sigma_n\}_{n\in\underline{N}}$;

(iii) P is a finite set of <u>productions</u> of the form $A(x_1,\ldots,x_n) \to t$, where $A \in N_n$ and $t \in W_V(X_n)$;

(iv) S is called the <u>start symbol</u> or <u>axiom</u> of G and $S \in N_0$.

Given s, $s' \in W_\Sigma(X_n)$ and $G = (N,\Sigma,P,S)$, s is said to <u>directly derive</u> s' (denoted by $s \underset{G}{\Rightarrow} s'$) if and only if s' is obtained from s by replacing <u>one</u> sub-expression of s of the form $At_1\ldots t_n$ by the expression $t \circ \langle t_1,\ldots,t_n\rangle$ where $A(x_1,\ldots,x_n) \to t$ is a production of G. Denote by $\overset{*}{\underset{G}{\Rightarrow}}$ the reflexive, transitive closure of $\underset{G}{\Rightarrow}$. Note that we will often drop the G from $\underset{G}{\Rightarrow}$ or $\overset{*}{\underset{G}{\Rightarrow}}$ whenever it is clear from the context which grammar G is being referred to.

A grammar $G = (N,\Sigma,P,S)$ is said to be <u>regular</u> if $N_n = \phi$ for all $n > 0$.

The set $L(G) = \{t \in W_\Sigma | S \overset{*}{\Rightarrow} t\}$ is called the (<u>term</u>) <u>language generated</u> by G. The language generated by a context free (regular) grammar $G = (N,\Sigma,P,S)$ is said to be a context free (regular) language (over Σ).

A direct derivation $s \Rightarrow s'$ in a grammar G is said to be leftmost if the non-terminal A in the subexpression $At_1 \ldots t_n$ of s which is to be replaced is the leftmost non-terminal symbol in s (regarded as a string of symbols). A derivation $S \stackrel{*}{\Rightarrow} s$ is leftmost if each step is leftmost.

Theorem: Let G be a context free term grammar. If $t \in L(G)$, then t has a leftmost derivation in G. $\square$

Let $CF_\Sigma = \{L \mid L = L(G)$, G a CF grammar over $\Sigma\}$, $REG_\Sigma = \{L \mid L = L(G)$, G a regular grammar over $\Sigma\}$.

A context free grammar $G = (N,\Sigma,P,S)$ is said to be in (Chomsky) normal form if each production in P is in one of the following forms:

(i) $A(x_1,\ldots,x_n) \to B(C_1(x_1,\ldots,x_n),\ldots,C_m(x_1,\ldots,x_n))$;

(ii) $A(x_1,\ldots,x_n) \to fx_1 \ldots x_n$

(iii) $A(x_1,\ldots,x_n) \to x_k$

for $A,C_1,\ldots,C_m \in N_n$, $B \in N_m$, $f \in \Sigma_n$, and $1 \le k \le n$ and if G has no useless non-terminals. A non-terminal $A \in N_n$ in a grammar $G = (N,\Sigma,P,S)$ is said to be useless if either:

(i) $\{t \mid t \in W_\Sigma(X_n)$ and $A(x_1,\ldots,x_n) \stackrel{*}{\Rightarrow} t\} = \phi$

or (ii) A is never used in a derivation in G starting at S.

See Rounds for further details.

Theorem (Maibaum): Given a context free term grammar G, there (effectively) exists a grammar G' in normal form such that $L(G) = L(G')$. $\square$

Theorem (Brainerd): Given a regular term grammar G, there (effectively) exists a regular term grammar G' such that L(G) = L(G') and G' only has productions of the form $A \to fB_1 \ldots B_n$ or $A \to a$ for non-terminals $A, B_1, \ldots, B_n$ and terminals f (of rank n) and a (of rank 0). □

The depth of an expression $t \in W_\Sigma(X)$, denoted by $|t|$, is defined as follows:

(i) $|t| = 0$ if $t = x$, $x \in X$;

(ii) If $t = ft_1 \ldots t_n$, then $|t| = 1 + \max\{|t_i|\}$.

If $\alpha \in B(W_\Sigma(X_n))$, then $|\alpha| = \max\{|t| \mid t \in \alpha\}$.

If $\alpha = \langle \alpha_1, \ldots, \alpha_n \rangle \in (B(W_\Sigma(X_m)))^n$, then $|\alpha| = \max\{|\alpha_i| \mid 1 \le i \le n\}$.

2. We use the preceding definitions to present pumping lemmas for regular and context free term grammars.

Theorem: Given a regular set L over Σ, there exists a constant $r > 0$ (depending only on L) such that, if $t \in L$ and $|t| > r$, then t can be written as $u_1 \circ u_2 \circ u_3$ where:

(i) $u_1(x_1!) \in W_\Sigma(X_1)$;

(ii) $u_2(x_1!) \in W_\Sigma(X_1)$ and $1 \le |u_2| \le r$;

(iii) $u_3 \in W_\Sigma$.

Moreover, $u_1 \circ u_2^i \circ u_3 \in L$ for all $i \ge 0$, where u_2^i is defined by:

(i) $u_2^0 = x_1$

(ii) $u_2^{i+1} = u_2^i \circ u_2$.

Proof: Let $L = L(G)$ where $G = <N,\Sigma,P,S>$ is a regular grammar with productions only of the form $A \to fB_1 \ldots B_m$ or $A \to a$. (Recall $N_k = \phi$ for $k > 0$.) Let $N_0 = \{A_1,\ldots,A_n\}$ and $r = n$. If $t \in L$ such that $|t| > r$, then we claim that there exists $A_j \in N_0$ and a derivation for t in G of the form:

$$\begin{aligned} S &\overset{*}{\Rightarrow} u_1 \circ A_j \\ &\overset{*}{\Rightarrow} u_1 \circ u_2 \circ A_j \\ &\overset{*}{\Rightarrow} u_1 \circ u_2 \circ u_3 \end{aligned}$$

for u_1,u_2,u_3 as in the statement of the theorem. If this were not so (i.e. no such A_j existed), then for each $A \in N_0$ we would have the following simple property: No derivation of the form $A \overset{*}{\Rightarrow} u \circ A$ exists for any $u(x_1!) \in W_\Sigma(X_1)$. That is, for all $t' \in W_\Sigma$ such that $A \overset{*}{\Rightarrow} t'$, we would have $|t'| \le n$. In particular, if $t' \in L$, then $|t'| \le n$. This contradicts our assumption that $|t| > n$.

But, then the following derivation is also a derivation in G for each $i \ge 0$:

$$\begin{aligned} S &\overset{*}{\Rightarrow} u_1 \circ A_j \\ &\overset{*}{\Rightarrow} u_1 \circ u_2 \circ A_j \\ &\overset{*}{\Rightarrow} u_1 \circ u_2 \circ u_2 \circ A_j \\ &\;\;\vdots \\ &\overset{*}{\Rightarrow} u_1 \circ \underbrace{u_2 \ \ldots\ u_2}_{\text{i-occurences}} \circ A_j \\ &\overset{*}{\Rightarrow} u_1 \circ u_2^i \circ u_3. \end{aligned}$$

□

Corollary: The emptiness and finiteness problems are solvable for regular term grammars. (See also Thatcher and Wright).

Proof If there is $t \in L(G)$, then, by the pumping lemma above, there is $t' \in L(G)$ such that $|t'| \le r$. Thus to test if $L(G)$ is empty, we need to check for the existence of such a t'. That this can be done follows from the following facts:

(i) The number of terms in W_Σ of depth less than or equal to r is finite;

(ii) Given $k \ge 0$ and G, the length of derivations $S \overset{*}{\Rightarrow} t$ for $|t| = k$ is bounded. (The length of a derivation is the number of rules applied in the course of the derivation.)

As to the finiteness problem, it is clear that if there is $t \in L(G)$ such that $r < |t| \le 2r$, then, by the pumping lemma above $L(G)$ is infinite. Conversely, if $L(G)$ is infinite, then there is $t \in L(G)$ such that $|t| > 2r$. (This follows from fact (i) above.) But then, applying the pumping lemma above, we can produce $t' \in L(G)$ such that $r < |t'| \le 2r$. Thus $L(G)$ is infinite if and only if there is $t \in L(G)$ such that $r < |t| \le 2r$. Thus, based on facts (i) and (ii) above, given $L(G)$, we can test for the existence of such a t. □

The pumping lemma for context free term grammars reads as follows:

Theorem: Given a context free language L over Σ, there exists constants $p,q > 0$ such that, if $t \in L$ and $|t| > p$, then $t \in u_1(u_2(u_5,u_3(u_4(u_5)))) = u_1(u_2(x_1,\ldots,x_n,u_3(u_4(x_1,\ldots,x_n)))\circ u_5)$ where:

(i) $u_1(x_1!) \in W_\Sigma(X_1)$;

(ii) $u_2(x_1,\ldots,x_n,x_{n+1}!) \in W_\Sigma(X_{n+1})$;

(iii) $u_3(x_1,\ldots,x_n) \in W_\Sigma(X_n)$;

(iv) $u_4(x_1,\ldots,x_n) \in (B(W_\Sigma(X_n)))^n$;

(v) $u_5 \in (B(W_\Sigma))^n$.

Moreover, if $t' \in u_2(x_1,\ldots,x_n,u_3(u_4(x_1,\ldots,x_n)))$, then $|t'| \le q, |u_2| + |u_4| > 0$ and, if we define

$$\theta^0 = u_3(x_1,\ldots,x_n)$$

and

$$\theta^{i+1} = u_2(x_1,\ldots,x_n,\theta^i(u_4(x_1,\ldots,x_n))),$$

then we have $u_1(\theta^i(u_5)) \subseteq L$ for all $i \ge 0$. (Note that $t \in u_1(\theta^1(u_5))$.) Before we proceed with the proof, we give below an intuitive outline along with some technical details formalising part of this intuition:

In the case of context free string grammars, we are able to prove the usual pumping lemma because of the existence of derivations of the form

$$S \overset{*}{\Rightarrow} uAs \overset{*}{\Rightarrow} uvArs \overset{*}{\Rightarrow} uvwrs$$

for strings of terminals u,v,w,r,s and non-terminal symbols S (the axiom) and A. That such derivations exist can be shown by studying derivations, or more precisely, derivation trees larger than some specified size.

In the case of context free term grammars, we must look for derivations of the form

$$\begin{aligned} S &\overset{*}{\Rightarrow} u(A(x_1,\ldots,x_n)\circ s) \\ &\overset{*}{\Rightarrow} u(v(x_1,\ldots,x_n,A(x_1,\ldots,x_n)\circ r(x_1,\ldots,x_n))\circ s) \\ &\overset{*}{\Rightarrow} u(v(x_1,\ldots,x_n,w(x_1,\ldots,x_n)\circ r(x_1,\ldots,x_n))\circ s) = t \end{aligned}$$

for trees of terminal symbols $u(x_1!)$, $v(x_1,\ldots,x_n,x_{n+1}!)$, $w(x_1,\ldots,x_n)$; n-tuples of trees of terminal symbols s and $r(x_1,\ldots,x_n)$; and non-terminals S and A. How can we guarantee the existence of such derivations?

First of all, there is no readily available concept of a derivation tree for a derivation in a context free term grammar. Thus the route followed in the case of string grammars is not readily available to us.

Consider, however, the following "analysis". The first indicated appearance of A in the above derivation appears "in place of" x_1 in $u(x_1)$. That is, there is a path from the root of $u_1(x_1)$ to x_1, call it p_1x_1, so that any path in $u_1(A(x_1,\ldots,x_n)\circ s)$ through A is p_1Ap for some p. Consider the second appearance of A in the above derivation. This second occurrence of A occurs "in place of" x_{n+1} in $u(v(x_1,\ldots,x_n,x_{n+1}!))$. That is, there is a path through this term, call it $p_1p_2x_{n+1}$, so that any path in $u(v(x_1,\ldots,x_nA(x_1,\ldots,x_n)\circ r(x_1,\ldots,x_n))\circ s)$ through A is p_1p_2Ap' for some p'. Now consider the result of the above derivation, $u(v(x_1,\ldots,x_n, w(x_1,\ldots,x_n)$ $\circ r(x_1,\ldots,x_n))\circ s)$. There must exist a path p_1p_2p'' through this term so that p" begins with the symbol labelling the root of $(w(x_1,\ldots,x_n)\circ r(x_1,\ldots,x_n))\circ s$.

Now, the two occurrences of A in the above derivation have made some contribution to the nature of the string p_1p_2p''. Conversely, the properties of paths through a term such as t can aid us in finding derivations of the appropriate kind in the term grammar. In the sequel, we make precise the relationship between a context free term language and the string language made up of all the paths through the terms in the term language. (This analysis is based on the work of Rounds.)

Let Δ be a ranked alphabet and X any set. Let $\bar{\Delta}_n = \Delta_n \times \{1,\ldots,n\}$ and write f_i for $<f,i> \in \bar{\Delta}_n$. Let $\bar{\Delta} = \Delta_0 \cup (\bigcup_{n>0} \bar{\Delta}_n)$. Let λ be a symbol such that $\lambda \notin (\cup\Delta_n) \cup X$. For each $\sigma \in \Delta_0 \cup X$, define the set of σ-paths through $t \in W_\Delta(X)$ as follows:

$$P_\sigma(t) = \begin{cases} \phi \text{ if } t,\sigma \in \Delta_0 \cup X \text{ and } t \neq \sigma \\ \{\sigma\} \text{ if } t = \sigma \text{ and } \sigma \in \Delta_0 \\ \{\lambda\} \text{ if } t = \sigma \text{ and } \sigma \in X \\ \bigcup_{i=1}^{n} \{ f_i w | w \in P_\sigma(t_i)\} \text{ if } t = ft_1 \ldots t_n. \end{cases}$$

For $L \subseteq W_\Delta(X)$, let $P(L) = \bigcup_{t \in L} (\bigcup\{P_\sigma(t) | \sigma \in \Delta_0 \cup X\})$

Let $G = \langle N,\Sigma,P,S\rangle$ be a context free term grammar. We construct a context free string grammar $\bar{G} = \langle\bar{N},\bar{\Sigma},\bar{P},\bar{S}\rangle$ (which will have the property that $L(\bar{G}) = P(L(G)))$ as follows:

(i) $\bar{N}$ and $\bar{\Sigma}$ are defined as above;

(ii) $\bar{S} = S \in N_0$;

(iii) $\bar{P}$ is obtained from the productions $A(x_1,\ldots,x_n) \to t$ in P as follows:

(a) If $w\lambda \in P_{x_i}(t)$, let $A_i \to w$ be in $\bar{P}$;

(b) If $\lambda \in P_{x_i}(t)$, let $A_i \to e$ (where e is the empty string) be in $\bar{P}$;

(c) If $wa \in P_a(t)$ for $a \in \Sigma_0$, let $A_i \to wa$ be in $\bar{P}$.

<u>Lemma (Rounds)</u>: If $L = L(G)$ is a context free term language, then $P(L)$ is a context free set of strings and $P(L) = L(\bar{G})$ with $\bar{G}$ as defined above. □

<u>Remark</u>: If G above is in normal form, then all productions in $\bar{G}$ are in one of the following forms:

(i) $A_i \to B_jC_k$ for some $A \in N_m$ $(1 \le i \le m)$, $B \in N_n$ $(1 \le j \le n)$, and $C \in N_p (1 \le k \le p)$;

(ii) $A_i \to a$ for some A in N and $a \in \Sigma_0$ or $a = f_i$ for $f \in \Sigma_m$;

(iii) $A_i \to e$ for some A in N.

Moreover, a given (leftmost) derivation $d = S \overset{*}{=>} t$, $t \in W_\Sigma$, in G induces a corresponding set of (leftmost) derivations $\mathcal{D}_d = \{S \overset{*}{=>} w | w \in P(\{t\})\}$ in $\bar{G}$. It is easily seen how this can be done: Suppose we have a derivation $S \overset{*}{=>} s => s'$ in G and we are given the set $\mathcal{D}=\mathcal{D}_{S \overset{*}{=>} s}$. Assume that s' is obtained from s by replacing a subexpression $At_1 \ldots t_n$ of s by $t' \circ \langle t_1,\ldots,t_n \rangle$ for some $A(x_1,\ldots,x_n) \to t'$ in P. Let $v\ A_i\ v'$ be some path in s through this non-terminal A. Depending on the form of t', we obtain $\mathcal{D}' = \mathcal{D}_{S \overset{*}{=>} s'}$ as follows:

(i) If $t' = x_k$ then we have two cases:

(a) If $i = k$, place $S \overset{*}{=>} v\ A_k\ v' => vv'$ (using $A_k \to e$ in $\bar{P}$) in $\mathcal{D}'$ for $S \overset{*}{=>} v\ A_k\ v'$ in $\mathcal{D}$.

(b) If $i \neq k$, then $S \overset{*}{=>} v\ A_i\ v'$ in $\mathcal{D}$ is <u>not</u> replaced in $\mathcal{D}'$.

(ii) If $t' = fx_1 \ldots x_n$, place $S \overset{*}{=>} vA_kv' => vf_kv'$ (using $A_k \to f_k$ in $\bar{P}$) in $\mathcal{D}'$ for $S \overset{*}{=>} vA_kv'$ in $\mathcal{D}$.

(iii) If $t' = B(C_1(x_1,\ldots,x_n),\ldots,C_m(x_1,\ldots,x_n))$, place the set $S \overset{*}{=>} vA_kv' => vB_jC_{j,k}v' | 1 \leq j \leq m\}$ (using $A_k \to B_jC_{j,k} (1 \leq j \leq m)$ in $\bar{P}$) in $\mathcal{D}'$ for $S \overset{*}{=>} vA_kv'$ in $\mathcal{D}$. $\square$

<u>Proof of the theorem</u>:

Let $L = L(G)$ where $G = \langle N,\Sigma,P,S \rangle$ is a normal form grammar. Suppose there are m non-terminals in N. Let $p = 2^{m-1}$ and $q = 2^m$. Let $d = S \overset{*}{=>} t$ be a leftmost derivation in G such that $t \in L$ and $|t| > p$. Let w be a path of maximum length in t. (Then the length of $w > 2^{m-1}$.) Thus there is a leftmost derivation $S \overset{*}{=>} w$ in $\mathcal{D}_d$. Construct the derivation tree T_w corresponding to this derivation in $\bar{G}$.

We will call a node of T_w labelled by some non-terminal $A_i \in \bar{N}$ <u>productive</u> if:

(i) A_i has as direct descendents the non-terminals B_j and C_k (i.e. we have used the production $A_i \to B_jC_k$);

and (ii) the two sets of terminal symbols labelling leaves which are descendents of B_j and C_k, respectively, are non-empty.

Condition (ii) implies that both B_j and C_k "contribute" non-empty substrings to w. It is a simple exercise to prove that there is some path ω in T_w which contains at least m productive nodes. (Choose ω so that the number of productive nodes on ω is maximised.) (This is possible since the length of w is greater than 2^{m-1} and so the depth of T_w is at least m.) But then ω must have at least m+1 nodes labelled by non-terminals in $\bar{N}$. (The last non-terminal in any path in T_w must be non-productive.)

This then implies that there is some A in N such that:

A: (i) Two of the nodes in ω are labelled by A_i and A_j for some $1 \le i, j \le n$ where $A \in N_n$;

(ii) A_i appears in ω before A_j;

(iii) A_i is productive and the number of productive nodes in ω which appear after A_i is at most m-1.

Condition A(i) can be met since there are only m non-terminals in N. Conditions A(ii) and A(iii) can be met by choosing the least postfix of ω containing m productive nodes.

Since A_i appears in ω, the leftmost derivation d can be expressed as

$$S \overset{*}{=>} u_1' \; (A(x_1,\ldots,x_n) \circ u_5') \overset{*}{=>} t$$

for some $u_1' \in W_V(\{x_1!\})$ and $u_5' = \langle t_1,\ldots,t_n\rangle \in (W_V)^n$ such that:

B: (i) A is the leftmost non-terminal in $u_1'(A(x_1,\ldots x_n) \circ u_5')$,

(ii) $u_1' \overset{*}{=>} u_1$ for some $u_1 \in W\;(\{x_1!\})$.

(Because of condition B(i), the unique occurrence of x_1 in u_1' does not appear to the right of any non-terminal and so will not be copied or dropped, fulfilling condition B(ii).)

Since A_j appears in ω after A_i, the leftmost derivation d can be expressed as:

$$\begin{aligned} S &\overset{*}{=>} u_1' \; (A(x_1,\ldots,x_n) \circ u_5') \\ &\overset{*}{=>} t',\; t' \in u_1'(u_2'(x_1,\ldots,x_n,A(x_1,\ldots,x_n) \circ u_4'(x_1,\ldots,x_n)) \circ u_5'') \\ &\overset{*}{=>} t \end{aligned}$$

for some $u_2' \in W_V(\{x_1,\ldots,x_n,x_{n+1}!\})$, $u_4' = \langle s_1,\ldots,s_n\rangle \in (W_V(X_n))^n$ and $u_5'' = \langle\tilde{t}_1,\ldots,\tilde{t}_n\rangle \in (B(W_V))^n$ such that:

C: (i) A is the leftmost non-terminal in $u_2'(x_1,\ldots,x_n, A(x_1,\ldots,x_n)\circ u_4'(x_1,\ldots,x_n))$:

(ii) $u_2' \overset{*}{\Rightarrow} u_2$ for some $u_2 \in W_\Sigma(\{x_1,\ldots,x_n,x_{n+1}!\})$;

(iii) If $r \in \tilde{t}_j$, $1 \le j \le n$, then $t_j \overset{*}{\Rightarrow} r$ for t_j in $u_5' = \langle t_1,\ldots,t_n\rangle$.

(Conditions C(i) and C(ii) can be explained as in the above paragraph. Condition C(iii) is justified as follows: t_j in u_5' can be substituted in a number of different places for x_j in u_2'. Each of these copies can lead to separate derivations from t_j.)

We can then write d as

$$S \overset{*}{\Rightarrow} u_1'(A(x_1,\ldots,x_n)\circ u_5')$$

$$\overset{*}{\Rightarrow} t',\ t' \in u_1'\ (u_2'(x_1,\ldots,x_n,A(x_1,\ldots,x_n)\circ u_4'(x_1,\ldots,x_n))\circ u_5'')$$

$$\overset{*}{\Rightarrow} t,\ t \in u_1(u_2(x_1,\ldots,x_n),u_3(x_1,\ldots,x_n)\circ u_4(x_1,\ldots,x_n)\circ u_5)$$

for some $u_3 \in W_\Sigma\ (X_n)$, $u_4 = \langle\tilde{s}_1,\ldots,\tilde{s}_n\rangle \in (B(W_\Sigma(X_n)))^n$ and $u_5 = \langle\tilde{t}_1',\ldots,\tilde{t}_n'\rangle \in (B(W_\Sigma(X_n)))^n$ such that:

D: (i) If $r \in \tilde{s}_j$, $1 \le j \le n$, then $s_j \overset{*}{\Rightarrow} r$ for s_j in u_4';

(ii) If $r \in \tilde{t}_j'$, $1 \le j \le n$, then $r' \overset{*}{\Rightarrow} r$ for some $r' \in \tilde{t}_j(\tilde{t}_j$ in $u_5'')$.

(The justification of conditions D(i) and D(ii) is similar to that for condition C(iii) in the above paragraph.)

To summarise, we have $u_1,\ldots,u_5$, as in the statement of the theorem, such that $S \overset{*}{\Rightarrow} r$, $r \in u_1(A(x_1,\ldots,x_n)\circ u_5)$

$$A(x_1,\ldots,x_n) \overset{*}{\Rightarrow} r', r' \in u_2(x_1,\ldots,x_n,\ A(x_1,\ldots,x_n)\circ\ u_4(x_1,\ldots,x_n))$$

and

$$A(x_1,\ldots,x_n) \overset{*}{\Rightarrow} u_3(x_1,\ldots,x_n).$$

We now proceed to prove the other claims made in the statement of the theorem. Since A_i was chosen to be productive, we can be sure that $|u_2|+|u_4| > 0$. Because of the way w, ω and A_i,A_j were chosen, we can be sure that if $t' \in u_2(x_1,\ldots,x_n,u_3(x_1,\ldots,x_n)\circ u_4(x_1,\ldots,x_n))$, then $|t'| \leq q$. Moreover, it is clear that using the derivations $A(x_1,\ldots,x_n) \overset{*}{\Rightarrow} t'$, $t' \in u_2(x_1,\ldots,x_n, A(x_1,\ldots,x_n)\circ u_4(x_1,\ldots,x_n))$, iteratively, we can produce the derivations $A(x_1,\ldots,x_n) \overset{*}{\Rightarrow} s'$, $s' \in \theta^i$ for any $i \geq 0$. Thus we have the last condition of the theorem: If $t \in u_1(\theta^i(u_5))$, then $S \overset{*}{\Rightarrow} t$.

Corollary: The emptiness and finiteness problems are solvable for context free term grammars.

Proof: It is clear from the theorem that if $t \in L(G)$ and $|t| > p$ (p as in the theorem), then we can produce another $t' \in L(G)$ such that $t' \in L(G)$ and $|t'| \leq p$. (We may need many applications of the theorem.) Thus to check whether L(G) is empty, it is sufficient to check whether there is some t, $|t| \leq p$, in L(G). That this can be done is clear from the following facts:

(i) The number of terms in W_Σ of depth less than or equal to p is finite;

(ii) Given $k \geq 0$ and G, the lengths of derivations $S \overset{*}{\Rightarrow} t$ for $|t| = k$ is bounded.

As to the finiteness problem, it is clear that if there is a $t \in L(G)$ such that $p < |t| \leq p + q$, then, by the pumping lemma above, L(G) is infinite. Conversely if L(G) is infinite, then there is $t \in L(G)$ such that $|t| > p + q$. (This follows from fact (i) above.) But then, by applying the pumping lemma above several times, we can produce $t' \in L(G)$ such that $p < |t'| \leq p + q$. Thus L(G) is infinite if and only if there exists $t \in L(G)$ such that $p < |t| \leq p + q$.

Thus, based on facts (i) and (ii) above, given L(G), we can test for the existence of such a t.

(Note that p and q, above depend on the number of non-terminals in G.)

Corollary: The emptiness problem is solvable for indexed grammars.

Proof: This is a simple consequence of the relationship between indexed languages and context free languages. For the definition of indexed languages and grammars, see Aho. For the connection between indexed languages and context free term languages, see Rounds or Maibaum. □

3. Let $\Sigma_0 = \{a\}$, $\Sigma_2 = \{t\}$ and $\Sigma_n = \phi$ for $n \neq 0,2$. Consider the set L = {+aa,++aa+aa,+++aa+aa++aa+aa,...} over Σ. L is the set of balanced binary "trees" over a and + with interior nodes labelled by + and leaves (or exterior nodes) labelled by a.

<u>Lemma</u> The set L described above is not regular.

<u>Proof</u> Suppose L is regular. Then, by the pumping lemma, there exists a constant $r > 0$ such that, if $t \in L$ and $|t| > r$, then t can be written as $u_1 \circ u_2 \circ u_3$ with $1 \leq |u_2| \leq r$. Moreover, $u_1 \circ u_2^i \circ u_3 \in L$ for all $i \geq 0$. Note that $t' \in L$ has the property that all paths from the root of t' to any leaf of t' are of equal length. This is certainly not true of $u_1 \circ u_2^2 \circ u_3$. This is a contradiction. Thus, L is not regular. (In fact, it is context free.) □

Let L' = {+aa,++aa+aa,++++aa+aa++aa+aa+++aa+aa++aa+aa,...}. L' is a language over Σ and L' is the set of balanced binary trees (over + and a) of depths 2^n+1 for $n \geq 0$.

<u>Lemma</u> The set L' described above is not context free.

<u>Proof:</u> Suppose L' is context free. Then, by the pumping lemma, there exist constants $p,q > 0$ such that, if $t \in L$ and $|t| > p$, then $t \in u_1(u_2(x_1, \ldots, x_n, u_3(u_4(x_1,\ldots,x_n))) \circ u_5)$ with $|u_2(x_1,\ldots,x_n, u_3(u_4(x_1,\ldots,x_n)))| \leq q$ and $|u_2|+|u_4| > 0$. Moreover, $u_1(\theta^i(u_5)) \subseteq L'$ for all $i \geq 0$. Let $|u_2|+|u_4| = k$. Then $|u_1(\theta^i(u_5))| \leq |t| + (i-1)k$ for $i>0$. That is, the depths of these terms in L' are bounded by an arithmetic progression $|t|$, $|t|+k$, $|t|+2k, \ldots$. The depths of terms in L', on the other hand, form a geometric progression $2,3,5,17,\ldots$, $|t| = 2^j+1$, $2^{j+1}+1$, $2^{j+2}+1, \ldots$. Thus the two series, starting from $|t|$, must differ at some point. This is a contradiction. Thus, L' is not context free. (In fact, it is an indexed

term language (Maibaum and Opatrný).) □

Acknowledgement: I am very grateful to the referee for his helpful comments and suggestions.

REFERENCES

ADJ (Goguen, J.A., Thatcher, J.W.,Wagner E.G., and Wright,J.B.),Initial Algebra Semantics and Continuous Algebras, JACM 24 (1977), 68-95.

Aho, A.V., Indexed grammars - an extension of context free grammars, JACM 15 (1968), 647-671.

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

Brainerd, W.E., Treegenerating regular systems, Inf. and Con. 14 (1969), 217-231.

Hayashi, T., On derivation trees of indexed grammars, Publ. RIMS Kyoto Univ. 9 (1973), 61-92.

Maibaum, T.S.E., A generalized approach to formal languages, J. Comput. Syst. Sci. 8 (1974), 409-439.

Maibaum, T.S.E., and Opatrný, J., Generalised indexed grammars and indexed term grammars, in preparation.

Moore, E.F., Gedanken experiments on sequential machines, Automata Studies, Princeton University Press, Princeton, N.J., 129-153, 1956.

Ogden, W.F., Intercalation theorems for stack automata, Proc. ACM Symp. on Theory of Computing, May 1968.

Rounds, W.C., Tree-oriented proofs of some theorems on context-free and indexed languages, Proc. 3rd ACM Symp. on Theory of Computing, May 1970.

Thatcher, J.W., Generalized^2 sequential machines, J. Comput. Syst. Sci. 4 (1970), 339-367.

Thatcher, J.W., Private communication, 1972.

Thatcher, J.W., and Wright, J.B., Generalized finite automata theory with an application to a decision problem of second-order logic, Math. Syst. Theory 2 (1968), 57-81.