PUMPING LEMMAS FOR TERM LANGUAGES

by

T.S.E. Maibaum

Research Report CS-75-12

Department of Computer Science

University of Waterloo
Waterloo, Ontario, Canada

April 1975

This research was supported by University of Waterloo
Research Grant No. 126-7024-02.

PUMPING LEMMAS FOR TERM LANGUAGES*

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

* This research was supported by University of Waterloo Research Grant No. 126-7024-02.

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).

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 sets. 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. We will now restrict our discussion to finite alphabets.

A Σ-algebra is a pair consisting of a set A, called the carrier 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 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 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 expressions or terms generated 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 $f_{W_\Sigma(X)}(t_1,\ldots,t_n) = ft_1 \ldots t_n$.

A homomorphism 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 $a_1,\ldots,a_n \in A$ and $f \in \Sigma_n$.

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

We now define the operation of substitution on the set $W_\Sigma(X_n)$, where $X_n = \{x_1,\ldots,x_n\}$. (See also Thatcher (1970), (1972) and Wagner.)

We will denote by $Sub_{X_n}(t;t_1,\ldots,t_n)$ the operation of simultaneously substituting (for $1 \le i \le n$) t_i for every occurrence of x_i in t. Note that if $t_1,\ldots,t_n \in W_\Sigma(X_m)$, then $Sub_{X_n}(\_;t_1,\ldots,t_n)$ is the unique homomorphism $\bar{\phi}:W_\Sigma(X_n) \to W_\Sigma(X_m)$ defined by the assignment $\phi(x_i) = t_i$, $1 \le i \le n$.

We will use the informal notation $t[t_1,\ldots,t_n]$ for the image of t under the homomorphism $Sub_{X_n}(\_;t_1,\ldots,t_n)$.

A <u>context free term grammar</u> (Rounds, Maibaum) G 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 $Sub_{X_n}(t;t_1,\ldots,t_n)$, 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 to which grammar we are referring.

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).

Theorem The class of languages generated by regular grammars is a proper subclass of the class of languages generated by context free grammars. □

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) $A(x_1,\ldots,x_n) \to fx_{j_1}\ldots x_{j_m}$;

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

for $A,C_1,\ldots,C_m \in N_n$, $B \in N_m$, $f \in \Sigma_m$, $1 \le j_i$, $k \le n$, and $1 \le i \le m$.

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

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 $|t| = 1 + \max_{1\le i\le n} \{|t_i|\}$.

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

Theorem Given a regular language 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 $u_1[u_2[u_3]]$ where:

(i) $u_1 \in W_\Sigma(\{y\})$ with exactly one occurrence of y;

(ii) $u_2 \in W_\Sigma(\{y\})$ with exactly one occurrence of y and $1 \leq |u_2| \leq r$;

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

Moreover, $u_1[u_2^i[u_3]] \in L$ for all $i \geq 0$, where u_2^i is defined by:

(i) $u_2^0 = y$

(ii) $u_2^{i+1} = u_2^i[u_2]$.

Proof Let $L = L(G)$, where $G = (N,\Sigma,P,S)$ is a regular term grammar. (Note $N_n = \phi$ for $n > 0$). Let $N_0 = \{A_1,\ldots,A_n\}$ and $r = n$. Consider $t \in L$ such that $|t| > r$. Then we must have

$$S \stackrel{*}{\Rightarrow} u_1[A_j]$$

$$\stackrel{*}{\Rightarrow} u_1[u_2[A_j]]$$

$$\stackrel{*}{\Rightarrow} u_1[u_2[u_3]]$$

for u_1, u_2, u_3 as in the statement of the theorem and $A_j \in N_0$. (If we regard t as a tree, this statement can be justified in greater detail as follows:
A path of maximum depth in t must have been generated by expanding $|t|$ non-terminals. Since $|t| > r$, there must have been a repetition of a non-terminal, say A_j, along this path.)

But, then $S \overset{*}{\Rightarrow} u_1[A_j]$

$$\overset{*}{\Rightarrow} u_1[u_2[A_j]]$$

$$\overset{*}{\Rightarrow} u_1[u_2[u_2[A_j]]]$$

$$\overset{*}{\Rightarrow} u_1[u_2[u_2[u_3]]]$$

is also a valid derivation. Clearly $u_1[u_2^i[u_3]] \in L$ for all $i \geq 0$ and u_2^i defined as in the statement of the theorem.

<u>Corollary</u> The emptiness and finiteness problems are solvable for regular term grammars. (See also Thatcher and Wright.)

<u>Proof</u> For the emptiness problem, it is clear that we only need to test terms of depth less than or equal to r, for a given grammar G. This can be done since L(G) is recursive. Similarly, for the finiteness problem, we need to test terms of depth greater than r but less than or equal to 2r for membership in L(G). A positive (negative) answer in either case provides a positive (negative) answer for the corresponding decision problem. □

In order to prove the pumping lemma for context free term languages, we need a result of Rounds concerning the "set of paths" of $t \in W_\Delta(X)$, where Δ is some ranked alphabet.

For each $f \in \Delta_n$, let f_i be a new symbol for $1 \leq i \leq n$. i.e. let $\bar{\Delta} = \{f_i | f \in \Delta_n$ for some n and $1 \leq i \leq n\}$. For each $a \in \Delta_0$, define the set of a-paths through $t \in W_\Delta(X)$ as follows:

(i) $P_a(x) = \phi, x \in X;$

(ii) $P_a(b) = \begin{cases} \phi, b \in \Delta_0, b \neq a \\ \{a\}, a \in \Delta_0, a = b; \end{cases}$

(iii) $P_a(ft_1 \ldots t_n) = \bigcup_{i=1}^{n} \{f_i w | w \in P_a(t_i)\}.$

Thus the set of a-paths of t is a set of strings (ending in the symbol a) over the string alphabet $\bar{\Delta}$.

For $L \subseteq W_\Delta$, define

$$P(L) = \bigcup_{a \in \Delta_0} \bigcup_{t \in L} P_a(t).$$

<u>Lemma (Rounds)</u> If L is a context free term language, then P(L) is a context free set of strings.

<u>Proof</u> Let $G = (N, \Sigma, P, S)$ be a normal form grammar such that $L(G) = L$. (Assume G has no useless productions (Rounds).) We will convert the productions of G into the productions of a context free string grammar $G' = (\bar{N}, \bar{\Sigma}, \bar{P}, \bar{S})$.

Consider the following definition of x-paths of $t \in W_V(X)$, where $x \in X$:

(i) $P_x(a) = \phi, a \in V_0;$

(ii) $P_x(y) = \begin{cases} \phi, y \neq x \\ x^*, y = x \text{ and } x^* \text{ is a \underline{new} variable;} \end{cases}$

(iii) $P_x(At_1 \ldots t_n) = \bigcup_{i=1}^{n} \{A_i w \mid w \in P_x(t_i), A_i \in \bar{V}\}.$

Now, if $A(x_1, \ldots, x_n) \to t$ is a production in P, consider $P_{x_i}(t)$, $1 \leq i \leq n$, and $P_a(t)$, $a \in V_0$. If $wx^* \in P_{x_i}(t)$, put $A_i \to w$ in $\bar{P}$. If $x^* \in P_{x_i}(t)$, put $A_i \to e$ (the empty string) in $\bar{P}$. If $wa \in P_a(t)$, put $A_i \to wa$ in $\bar{P}$.

Let $\bar{S} = S$. G' is then obviously a context free string grammar. (Note that it is not quite in Chomsky normal form. There are some productions of the form $A_i \to e$ in $\bar{P}$, where $A_i \neq \bar{S}$.) It can be shown by induction that $L(G') = P(L(G))$. □

Theorem Given a context free language L over Σ, there exist constants $p,q > 0$ (depending only on L) such that, if $t \in L$ and $|t| > p$, then t can be written as $u_1[u_2[u_3[u_4[u_5]]]]$ where:

(i) $u_1 \in W_\Sigma(\{y\})$ with exactly one occurrence of y;

(ii) $u_3 \in W_\Sigma(\{X_n\})$;

(iii) $u_4 = (u_{41},\dots,u_{4n}) \in (W_\Sigma(\{X_n\}))^n$ (i.e. an n-tuple of terms);

(iv) $u_5 = (u_{51},\dots,u_{5n}) \in (W_\Sigma)^n$ (i.e. an n-tuple of terms);

(v) $u_2 = \bar{u}_2[y,u_{51},\dots,u_{5n}]$ and $\bar{u}_2 \in W_\Sigma(\{y,x_1,\dots,x_n\})$ with exactly one occurrence of y.

Moreover, $|u_2[u_3[u_4]]| \leq q$, $|u_2| + (\max_{1\leq i\leq n} \{|u_{4i}|\}) > 0$, and $u_1[u_2^i[u_3[u_4^i[u_5]]]] \in L$ for all $i \geq 0$ where:

(i) $u_4^i = (u_{41}^i,\dots,u_{4n}^i)$ and for $1 \leq j \leq n$:

(a) $u_{4j}^0 = x_j$;

(b) $u_{4j}^{k+1} = u_{4j}^k[u_{41}^k,\dots,u_{4n}^k]$

and (ii) u_2^k is defined by:

(a) $u_2^0 = y$;

(b) $u_2^{k+1} = (u_2^k[u_2[u_{41}^k,\dots,u_{4n}^k]])[u_{51},\dots,u_{5n}]$.

Proof Let $L = L(G)$, where $G = (N,\Sigma,P,S)$ is a normal form grammar. Suppose there are k non-terminals in N. Let $p = 2^{k-1}$ and $q = 2^k$. Consider $t \in L$ such that $|t| > p$. Then in P(t) there is a string w of length greater than $p = 2^{k-1}$. Consider the derivation tree of w in G' (the grammar of the previous lemma). There is a path in this derivation tree such that more thank non-terminals of $\bar{N}$ appear on it. Two of these occurrences must be A_i and A_j for some A in N and some i,j. This is because there are only k distinct non-terminals in N.

Then, in the original term grammar G, it must be true that

$$\begin{aligned} S &\overset{*}{\Rightarrow} u_1[A(x_1,\ldots,x_n)[u_5]] \\ &\overset{*}{\Rightarrow} u_1[u_2[A(x_1,\ldots,x_n)[u_4[u_5]]]] \qquad \text{(I)} \\ &\overset{*}{\Rightarrow} u_1[u_2[u_3[u_4[u_5]]]]. \end{aligned}$$

(Note that $A(x_1,\ldots,x_n) \overset{*}{\Rightarrow} u_2[A(x_1,\ldots,x_n)[u_4]]$

$$\overset{*}{\Rightarrow} u_2[u_3[u_4]].)$$

Again returning to G', it is clear that the occurrences of A_i and A_j can be chosen in such a way that $|u_2[u_3[u_4]]| \leq q$ and it is obvious that $|u_2| + \max_{1\leq i\leq n} \{|u_{4i}|\} > 0$. Moreover, the "middle" steps of the derivation I can be repeated as often as desired and so $u_1[u_2^i[u_3[u_4^i[u_5]]]] \in L$ for all $i \geq 0$. □

<u>Corollary</u> The emptiness and finiteness problems are solvable for context free term grammars. (See also Rounds for indirect proofs of these results.)

<u>Proof</u> For the emptiness problem, it is clear that we only need to test terms of depth less than or equal to p, for a given grammar G. This can be done since L(G) is recursive. Similarly, for the finiteness problem, we need to test terms of depth greater than p but less than or equal to p+q for membership in L(G). A positive (negative) answer in either case provides a positive (negative) answer for the corresponding decision problem. □

3. Let $\Sigma_0 = \{a\}$, $\Sigma_2 = \{+\}$ 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.

Lemma The set L described above is not regular.

Proof 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 $u_1[u_2[u_3]]$ with $1 \leq |u_2| \leq r$. Moreover, $u_1[u_2^i[u_3]] \leq L$ 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 $u_1[u_2^2[u_3]]$. 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 for $n \geq 0$.

Lemma The set L' described above is not context free.

Proof 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 can be written as $u_1[u_2[u_3[u_4[u_5]]]]$ with $|u_2[u_3[u_4]]| \leq q$ and $|u_2| + \max\limits_{1 \leq i \leq n} \{|u_{4i}|\} > 0$. Moreover, $u_1[u_2^i[u_3[u_4^i[u_5]]]] \in L'$ for all $i \geq 0$. Let $|u_2| + \max\limits_{1 \leq i \leq n} \{|u_{4i}|\} = k$. Then $u_1[u_2^i[u_3[u_4^i[u_5]]]] = |t| + (i-1)k$ for $i > 0$. That is, the depths of these terms (which are supposed to be in L') form an arithmetic progression

$|t|$, $|t|+k$, $|t|+2k, \ldots$. The depths of terms in L', on the other hand, form a geometric progression $2,4,16,\ldots, |t| = 2^j, 2^{j+1}, 2^{j+2}, \ldots$. 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).) □

REFERENCES

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. Kommunikation-forsch 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 (1973), 409-439.

Maibaum, T.S.E., and Opatrny, J., Generalised indexed grammars and indexed term grammars, In preparation.

Moore, E.F., Gedanken experiments on sequential machines, Automata Studies, Princeton Univ. 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.