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.
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 $\mathbb{N}$ be the set of natural numbers. A **ranked alphabet** is a family of sets indexed by $\mathbb{N}$. We use the notation $\Sigma = \{\Sigma_n\}_{n \in \mathbb{N}}$ for ranked sets. If $f \in \Sigma_n$, $f$ is said to be of rank $n$. $\Sigma$ is said to be finite if the (disjoint) union of $\{\Sigma_n\}_{n \in \mathbb{N}}$ is finite. We will now restrict our discussion to finite alphabets.

A $\Sigma$-algebra is a pair consisting of a set $A$, called the **carrier** of the algebra, and an indexed family of assignments $\alpha = \{\alpha_n\}_{n \in \mathbb{N}}$ 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 $\Sigma$-algebra with carrier $A$ by $A_\Sigma$.

Let $X$ be any set and consider the set $W_\Sigma(X)$ defined by:

1. $X \subseteq W_\Sigma(X)$;
2. If $f \in \Sigma_n$ and $t_i \in W_\Sigma(X)$ for $1 \leq i \leq n$, 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 $\Sigma$-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 $\Sigma$-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 $\Sigma$-algebra $A_\Sigma$ and an assignment $\phi : X \to A$, there is exactly one extension of $\phi$ to a homomorphism $\tilde{\phi} : W_\Sigma(X) \to A_\Sigma$. 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 \( \text{Sub}_n (t; t_1, \ldots, t_n) \) the operation of simultaneously substituting (for \( 1 \leq i \leq n \)) \( t_i \) for every occurrence of \( x_i \) in \( t \). Note that if \( t_1, \ldots, t_n \in W_L(X_n) \), then \( \text{Sub}_n (\_ ; t_1, \ldots, t_n) \) is the unique homomorphism \( \phi: W_L(X_n) \to W_L(X_n) \) defined by the assignment \( \phi(x_i) = t_i \), \( 1 \leq i \leq n \).

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

A context free term grammar (Rounds, Maibaum) \( G \) is a 4-tuple \((N, \Sigma, P, S)\) where:

(i) \( N \) is a finite ranked alphabet called the set of non-terminals of \( G \);

(ii) \( \Sigma \) is a finite ranked alphabet called the set of terminals of \( G \).

Let \( V = \{ V_n \}_{n \in \mathbb{N}} = \{ N_n \cup \Sigma_n \}_{n \in \mathbb{N}} \);

(iii) \( P \) is a finite set of productions 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 start symbol or axiom of \( G \) and \( S \in N_0 \).

Given \( s, s' \in W_L(X_n) \) and \( G = (N, \Sigma, P, S) \), \( s \) is said to directly derive \( s' \) (denoted by \( s \xrightarrow{G} s' \)) if and only if \( s' \) is obtained from \( s \) by replacing one sub-expression of \( s \) of the form \( A t_1 \ldots t_n \) by the expression \( \text{Sub}_n (t; t_1, \ldots, t_n) \), where \( A(x_1, \ldots, x_n) \to t \) is a production of \( G \). Denote by \( \xrightarrow{G}^* \) the reflexive, transitive closure of \( \xrightarrow{G} \). Note that we will often drop the \( G \) from \( \xrightarrow{G} \) or \( \xrightarrow{G}^* \) whenever it is clear to which grammar we are referring.

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

The set \( L(G) = \{ t \in W_L \mid S \xrightarrow{G}^* t \} \) is called the (term) language generated 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 \( \Sigma \)).
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) $A(x_1, \ldots, x_n) \rightarrow B(C_1(x_1, \ldots, x_n), \ldots, C_m(x_1, \ldots, x_n))$
(ii) $A(x_1, \ldots, x_n) \rightarrow fx_1 \ldots x_m$
(iii) $A(x_1, \ldots, x_n) \rightarrow x_k$

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

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 $t \in W_n(X)$, 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 $|t| = 1 + \max_{1 \leq i \leq n} \{ |t_i| \}$. 


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 $\Sigma$, 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

\[
S \Rightarrow u_1[A_j] \\
\Rightarrow u_1[u_2[A_j]] \\
\Rightarrow u_1[u_2[u_3]]
\]

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 \( S \Rightarrow u_1[A_j] \)
\[ \Rightarrow u_1[u_2[A_j]] \]
\[ \Rightarrow u_1[u_2[u_2[A_j]]] \]
\[ \Rightarrow u_1[u_2[u_2[u_3]]] \]

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.

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

**Proof** 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. \( \square \)

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 \( \Delta \) 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.
\[ \Delta = \{ f_i | f \in \Delta_n \text{ for some } n \text{ and } 1 \leq i \leq n \}. \]
For each \( a \in \Delta_0 \), define the set of \( a \)-paths through \( t \in W_\Delta(X) \) as follows:

1. \( P_a(x) = \emptyset, x \in X; \)
2. \( P_a(b) = \{ \emptyset, b \in \Delta_0, b \neq a \}
\{a\}, a \in \Delta_0, a = b; \)
3. \[ P_a(ft_1...t_n) = \bigcup_{i=1}^n \{ f_iw | w \in P_a(t_i) \}. \]
Thus the set of $a$-paths of $t$ is a set of strings (ending in the symbol $a$) over the string alphabet $\tilde{\Delta}$.

For $L \subseteq W_{\tilde{\Delta}}$, define

$$P(L) = \bigcup_{a \in \tilde{\Delta}_0} \bigcup_{t \in L} P_a(t).$$

**Lemma (Rounds)** If $L$ is a context free term language, then $P(L)$ is a context free set of strings.

**Proof** 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 $G' = (N, \tilde{\Delta}, \tilde{P}, \tilde{S})$.

Consider the following definition of $x$-paths of $t \in W_{\tilde{\Delta}}(X)$, where $x \in X$:

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

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

(iii) $P_x(A_{t_1} \ldots t_n) = \bigcup_{i=1}^{n} \{ A_i w | w \in P_{x_{t_i}}(A_i), A_i \in \tilde{V} \}$.

Now, if $A(x_1, \ldots, x_n) \rightarrow t$ 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 $A_i \rightarrow w$ in $\tilde{P}$. If $x^* \in P_{x_i}(t)$, put $A_i \rightarrow e$ (the empty string) in $\tilde{P}$. If $wa \in P_a(t)$, put $A_i \rightarrow wa$ in $\tilde{P}$.

Let $\tilde{S} = S$. $G'$ 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 $A_i \rightarrow e$ in $\tilde{P}$, where $A_i \neq \tilde{S}$.) It can be shown by induction that $L(G') = P(L(G))$. □
Theorem. Given a context free language \( L \) over \( \Sigma \), 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) \( u_1 \in W_{\Sigma}\left(\{y\}\right) \) with exactly one occurrence of \( y \);

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

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

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

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

Moreover, \(|u_2[u_3[u_4]]| \leq q, |u_2| + \left( \max \{ |u_{4i}| \} \right) > 0\), and \( u_1[u_2[u_3[u_4]][u_5]] \in L \) for all \( i \geq 0 \) where:

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

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

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

and (ii) \( u_2^k \) is defined by:

(a) \( u_2^0 = y \);

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

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 than \( k \) non-terminals of \( \tilde{N} \) 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
\[ S \Rightarrow u_1[A(x_1, \ldots, x_n)[u_5]] \]
\[ \Rightarrow u_1[u_2[A(x_1, \ldots, x_n)[u_4[u_5]]]] \]
\[ \Rightarrow u_1[u_2[u_3[u_4[u_5]]]]. \]  \hspace{1cm} (I)
(Note that $A(x_1, \ldots, x_n) \Rightarrow u_2[A(x_1, \ldots, x_n)[u_4]]$
\[ \Rightarrow u_2[u_3[u_4]]. \])

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
$|u_2| + \max \{|u_4_i|\} > 0$. Moreover, the "middle" steps of the derivation $I$ can
be repeated as often as desired and so $u_1[u_2[u_3[u_4[u_5]]]] \in L$ for all $i \geq 0$. \hfill \square

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

**Proof** 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. \hfill \square
3. Let $\Sigma_0 = \{a\}$, $\Sigma_2 = \{+\}$ and $\Sigma_n = \emptyset$ for $n \neq 0, 2$. Consider the set $L = \{+aa, ++aa+aa, +++aa+aa+aa, \ldots\}$ over $\Sigma$. $L$ 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, $u_1[u_2^i[u_3]] \in L$ 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 $u_1[u_2^2[u_3]]$. This is a contradiction. Thus, $L$ is not regular. (In fact, it is context free.) □

Let $L' = \{+aa, ++aa+aa, +++aa+aa+aa, \ldots\}$. $L'$ is a language over $\Sigma$ 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 $|u_2| + \max_{1 \leq i \leq n} |u_4_i| > 0$. Moreover, $u_1[u_2^i[u_3[u_4^i[u_5]]]] \in L'$ for all $i \geq 0$. Let $|u_2| + \max_{1 \leq i \leq n} |u_4_i| = k$. 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


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


