SUBWORDS IN DETERMINISTIC TOL LANGUAGES *

By

A. Ehrenfeucht **
G. Rosenberg ***

** Department of Computer Science
University of Colorado, Boulder, Colorado

*** Department of Computer Science
State University of New York at Buffalo

Report # CU-CS-015-73 March, 1973

* This work supported by N.S.F. Grant GJ-660.

1. INTRODUCTION.

Developmental systems are formal structures which model the way in which certain biological organisms develop. The study of such systems has recently attracted particular attention as a new branch of the theory of formal languages, see for example [1], [4], [6] and their references. In this paper we continue research into T0L systems which were introduced in [2] and further studies, for example, in [3], [5].

A T0L system has the following components

(i) A finite set of symbols, Σ, the alphabet,

(ii) A finite collection $\mathcal{P}$ of tables, each of which tell us by what strings in Σ^* a symbol may be replaced. A table may, in general, contain several productions for each symbol. In every step of a derivation, all symbols in a string must be simultaneously replaced according to the production rules of one arbitrarily chosen table.

(iii) A starting string, ω, the axiom. The language generated by a given T0L system consists of ω and all strings which can be derived from ω in a finite number of steps.

A T0L system is called deterministic (abbreviated a DT0L system) if each of its tables is such that for each symbol in the alphabet the table contains exactly one production with the symbol on the left. The role of a deterministic restriction is one of the important questions, from both the biological and formal points of view, in the theory of developmental systems.

In this paper we prove a result, which we believe is a fundamental one for the characterization of languages generated by DT0L systems (called *DT0L languages*.) This result says that if L is a DT0L language over an alphabet containing at least two letters then the relative number of subwords of a given length k occurring in the words of L tends to zero as k increases.

The formal definition of a T0L system is given for example in [3] the terminology and notation of which we shall follow in this paper. In addition to this we shall use the following notation:

1) If x is a word then $|x|$ denotes its length and if A is a set then $\# A$ denotes the cardinality of A. The empty word is denoted by Λ.

2) If $G = \langle \Sigma, \mathcal{P}, \omega \rangle$ is a DT0L system and $T \varepsilon \mathcal{P}$ then $a \underset{T}{\to} \alpha$ abbreviates "the production $a \to \alpha$ is in T", and for a word x $T(x)$ denotes the word y such that $x \underset{T}{\Rightarrow} y$.

3) If L is a language and k a natural number, then $P_k(L)$ denotes the set of all subwords of length k occurring in the words of L.

2. PRELIMINARY LEMMAS.

Let Σ be a finite alphabet where $\#\Sigma = n \geq 2$, and $Z = \{w_1,\dots,w_n\}$ a finite non-empty subset of Σ^* such that $|w| > 1$ for at least one w in Z.

First three lemmas state some properties of Z^*.

Lemma 1.

There exists k_o such that $P_{k_o}(Z^*) < n^{k_o}$.

Proof:

For a in Σ, let $W(Z,a) = \{w \varepsilon Z\text{: a occurs in } w\}$.

We can distinguish two cases.

Case I. For every a in Σ there exists a word w in $W(Z,a)$ such that $w = a^r$ for some $r \geq 0$.

Case II. There exists a letter a in Σ such that no word in $W(Z,a)$ is of the form a^r for some $r \geq 0$.

If Case I holds, then (because $\#Z = \#\Sigma$), for every a in Σ there exists exactly one w in Z such that $w = a^r$ for some $r \geq 0$. But then one of the words in Z is of the form b^ℓ for some $\ell > 1$, b in Σ, where b does not occur in any other word in Z. Consequently if $x \varepsilon \Sigma - \{b\}$ then $xbx \not\in Z^*$.

If Case II holds then we have to consider two subcases.

Subcase II.1. There exists a letter a in Σ such that $W(a,Z) = \emptyset$. Then $a \not\in Z^*$.

Subcase II.2. For every a in Σ, $W(a,Z) \neq \emptyset$. Then, obviously, $a^{2u_a} \not\in Z^*$ for an arbitrary a in Σ and $u_a = \max\{|w| : w \varepsilon W(Z,a)\}$.

Lemma 1 follows now from the above case analysis.

Lemma 2.

Let $k \varepsilon N^+$, $k = k_0 s + k_1$ where k_0 is a constant as in Lemma 1 and $k_1 < k_0$. Then

$$\frac{P_k(Z^*)}{n^k} \leq \left(\frac{P_{k_o}(Z^*)}{n^{k_o}}\right)^s .$$

Proof:

Let k, k_o, s, k_1 be as described above.

Obviously $P_k(Z^*) \leq (P_{k_o}(Z^*))^s \cdot n^{k_1}$.

Hence

$$\frac{P_k(Z^*)}{n^k} \leq \frac{(P_{k_o}(Z^*))^s \cdot n^{k_1}}{n^{k_o s + k_1}} = \left(\frac{P_{k_o}(Z^*)}{n^{k_o}}\right)^s$$

and Lemma 2 holds.

Lemma 3.

$$\lim_{k\to\infty} \frac{P_k(Z^*)}{n^k} = 0 .$$

Proof:

Let k_o be a constant as in Lemma 1, $k \varepsilon N^+$ and s be defined as in Lemma 2. Then from Lemma 2 it follows that $\frac{P_k(Z^*)}{n^k} \leq \left(\frac{P_{k_o}(Z^*)}{n^{k_o}}\right)^s$ and from Lemma 1 it follows that $\frac{P_{k_o}(Z^*)}{n^{k_o}} < 1$.

Hence

$$\lim_{k\to\infty} \frac{P_k(Z^*)}{n^k} \leq \lim_{s\to\infty} \left(\frac{P_{k_o}(Z^*)}{n^{k_o}}\right)^s = 0$$

and Lemma 3 holds.

For our next Lemma we shall need some additional notation. Let $G = <\Sigma, ,\omega>$ be a DT0L system. We shall now distinguish some subsets of $\mathcal{P}$ and accordingly some subsets of $L(G)$.

Let $\mathcal{P}_g = \{T \varepsilon \mathcal{P}$: $a \underset{T}{\rightarrow} \alpha$ for some a in Σ, α in Σ^* such that $|\alpha| \geq 2\}$,

$\mathcal{P}_c = \{T \varepsilon \mathcal{P}$: if $a \underset{T}{\rightarrow} \alpha$ then $|\alpha| = 1\}$, (each element of $\mathcal{P}_c$ is called a coding table),

$\overline{\mathcal{P}}_c = \{T \varepsilon \mathcal{P}$: if $a \underset{T}{\rightarrow} \alpha$ then $|\alpha| = 1$ and $b \underset{T}{\rightarrow} \Lambda$ for some b in $\Sigma\}$.

If $T \varepsilon \mathcal{P}$ then we define two languages "associated with T" as follows:

$L_T = \{x \varepsilon \Sigma^*$: there exists y in Σ^* such that $\omega \Rightarrow^* y \underset{T}{\Rightarrow} x\}$,

$\hat{L}_T = \{x \varepsilon \Sigma^*$: there exists y in L_T and $\overline{T}$ in $\mathcal{P}_c$ such that $y \underset{\overline{T}}{\Rightarrow} x\}$.

Now we shall show how an arbitrary DT0L language can be decomposed using this notation.

Lemma 4.

For every DT0L system G there exists an equivalent DT0L system H with a set of tables $\mathcal{P}$ such that

$$L(G) = F \cup \bigcup_{T \varepsilon \mathcal{P}_g} L_T \cup \bigcup_{T \varepsilon \mathcal{P}_g} \hat{L}_T \cup \bigcup_{T \varepsilon \overline{\mathcal{P}}_c} L_T$$

where F is a finite language.

Proof:

Let $G = \langle \Sigma, \mathcal{R}, \omega \rangle$ be a DT0L system.

We leave to the reader the easy proof of the fact that one may (effectively) construct a finite number of coding tables $P_1, P_2, \ldots, P_t$ (not necessarily in $\mathcal{R}$) such that for every x in Σ^* and every sequence $T_{i_1} T_{i_2} T_{i_r}$ of coding tables from $\mathcal{R} \cup \{P_1, \ldots, P_t\}$ there exists $j \varepsilon \{1, \ldots, t\}$ such that

$$T_{i_r} \ldots T_{i_2} T_{i_1}(x) = P_j(x).$$

Hence we have a DT0L system $H = \langle\Sigma, \mathcal{P}, \omega\rangle$ where $\mathcal{P} = (\mathcal{R}-\mathcal{R}_c) \cup \{P_1,\ldots,P_t\}$, which is equivalent to G. We leave to the reader the easy proof of the fact that if we set $F = \{\omega\} \cup \{P(\omega): P \varepsilon \mathcal{P}_c\}$ then indeed

$$L(G) = F \cup \bigcup_{T \varepsilon \mathcal{P}_g} L_T \cup \bigcup_{T \varepsilon \mathcal{P}_g} \hat{L}_T \cup \bigcup_{T \varepsilon \overline{\mathcal{P}}_c} L_T$$

and so Lemma 4 holds.

3. MAIN RESULT.

Theorem.

Let Σ be a finite alphabet such that $\#\Sigma = n \geq 2$.

If L is a DTOL language, $L \subseteq \Sigma^*$, then

$$\lim_{k\to\infty} \frac{P_k(L)}{n^k} = 0.$$

Proof:

Let Σ be a finite alphabet where $\#\Sigma = n \geq 2$ and L be a DTOL language ($L \subseteq \Sigma^*$) generated by a DTOL system $G = \langle \Sigma, \mathcal{P}, \omega \rangle$. According to Lemma 4 and its proof we may assume that

$$L = \mathcal{L}(G) = F \cup \bigcup_{T\varepsilon\mathcal{P}_g} L_T \cup \bigcup_{T\varepsilon\mathcal{P}_g} \hat{L}_T \cup \bigcup_{T\varepsilon\overline{\mathcal{P}}_c} L_T$$

where $F = \{\omega\} \cup \{T(\omega) : T\varepsilon\mathcal{P}_c\}$ is a finite language.

1) Obviously $\lim\limits_{k\to\infty} \frac{P_k(F)}{n^k} = 0$.

2) If $T\varepsilon\mathcal{P}_g$, then (as $L_T \subseteq \{\alpha : a \xrightarrow[T]{} \alpha\}^*$) from Lemma 3 it follows that $\lim\limits_{k\to\infty} \frac{P_k(L_T)}{n^k} = 0$ and consequently

$$\lim_{k\to\infty} \frac{P_k(\bigcup_{T\varepsilon_g} L_T)}{n^k} = 0.$$

3) Let $T\varepsilon\mathcal{P}_g$. If $\#\mathcal{P}_c = m$ then $\#\hat{L}_T \leq m\cdot(\#L_T)$ and consequently $\#P_k(\hat{L}_T) \leq m\cdot P_k(L_T)$. Thus from 2) it follows that $\lim\limits_{k\to\infty} \frac{P_k(\hat{L}_T)}{n^k} = 0$, hence

$$\lim_{k\to\infty} \frac{P_k(\bigcup_{T\varepsilon\mathcal{P}_g} \hat{L}_T)}{n^k}.$$

4) Let $T\varepsilon\overline{\mathcal{P}}_c$. There exists a letter in Σ which does not occur at the right hand side of any production in T. Thus

$P_k(L_T) \leq (n-1)^k$ and $\dfrac{P_k(L_T)}{n^k} \leq \dfrac{(n-1)^k}{n^k}$, hence $\lim\limits_{k\to\infty} \dfrac{P_k(L_T)}{n^k} = 0$ and consequently

$$\lim_{k\to\infty} \frac{P_k(\bigcup\limits_{T\varepsilon \overline{P}_c} L_T)}{n^k} = 0.$$

Thus from 1) through 4) and from the expression for $L(G)$ it follows that $\lim\limits_{k\to\infty} \dfrac{P_k(L)}{n^k} = 0$ and so the Theorem holds.

Up to the present, characterization results have been conspicuously absent from the theory, necessitating involved combinatorial proofs to show that certain languages are not DT0L. The result in this paper gives us a direct way to show that many languages are not generable by DT0L systems. For example if $\Sigma = \{a,b\}$ and F is a finite language over Σ then $\Sigma^* - F$ is not in the class of DT0L languages.

REFERENCES.

[1] A. Lindenmayer, G. Rozenberg, Developmental systems and languages, in Proc. IVth ACM Symp. on Theory of Comp., Denver, Colorado, 1972.

[2] G. Rozenberg, T0L systems and languages, in Information and Control, to appear.

[3] G. Rozenberg, The equivalence problem for deterministic T0L systems is undecidable, in Information Processing Letters, vol. 1, no. 5, 1972.

[4] G. Rozenberg, D0L sequences, in Discrete Mathematics, to appear.

[5] G. Rozenberg, Extension of tabled 0L systems and languages, submitted for publication.

[6] A. Salomaa, Formal Languages, Academic Press, 1973.