THE NUMBER OF OCCURRENCES OF LETTERS VERSUS THEIR DISTRIBUTION IN SOME EOL LANGUAGES.

A. Ehrenfeucht
Department of Computer Science
University of Colorado
Boulder, Colorado 80302

G. Rozenberg
Department of Mathematics
Utrecht University
Utrecht-Uithof, The Netherlands

Report #CU-CS-030-73 August, 1973

* This work is supported in part by National Science Foundation Grant #GJ-660

All correspondence to:

G. Rozenberg
Department of Mathematics
Utrecht University
Utrecht-Uithof
The Netherlands

ABSTRACT.

A characterization theorem is given for a class of developmental languages. The theorem binds together the number of occurrences of letters in the words of the given language with the distribution of these letters.

0. INTRODUCTION.

This paper deals with a class of developmental languages. The theory of developmental systems and languages originated in the works of Lindenmayer (see [Lindenmayer]). This theory provided a useful theoretical framework within which the nature of cellular behavior in development can be discussed, computed and compared (see, e.g., [Herman and Rozenberg], [Lindenmayer] and [Lindenmayer and Rozenberg]). It turned out that developmental systems and languages are interesting and novel objects from the formal language theory point of view. Especially in comparison with Chomsky grammars and languages (see, e.g., [Ginsburg]) they provided a lot of insight into the basic problems of formal language theory.

An important subclass of developmental systems are the so-called EOL systems (see, e.g., [Herman] or [Herman, Lindenmayer, and Rozenberg]) which were devised to allow descriptions of development which take into account the inaccuracy of our observations.

One of the basic open problems within the theory of EOL systems (and in fact within the whole theory of developmental systems) is the characterization theorems which allow one, for example, to prove that some languages are not EOL languages (i.e., languages generated by the EOL systems).

This paper provides such a characterization for a subclass of EOL languages. The characterization theorem binds together the number of occurrences (in the words of the given EOL language) of letters from a given set of letters with the distribution of these letters.

The paper also discusses some applications of the main result.

1. PRELIMINARIES.

We assume the reader to be familiar with the basics of formal language theory (see, e.g., [Ginsburg], whose notation and terminology we shall mostly follow). In addition to this, we shall use the following notation:

(i) N denotes the set of nonnegative integers and $N^+ = N - \{0\}$. If n is an integer, then abs(n) denotes its absolute value.

(ii) If x is a word over an alphabet Σ, then $|x|$ denotes the length of x and Min(x) denotes the set of letters which occur in x. For a in Σ, $\#_a(x)$ denotes the number of occurrences of the letter a in x and if B is a subset of Σ then $\#_B(x) = \sum_{a \varepsilon B} \#_a(x)$. If k is a positive integer, then x^k denotes x catenated k times with itself.

(iii) If A is a finite set, then #A denotes its cardinality. If $B \subseteq A$ and $\#B = 1$ then B is called a <u>singleton in</u> A.

(iv) A <u>coding</u> is a letter to letter homomorphism. If h is a homomorphism from Σ^* into V^* and $L \subseteq V^*$ then $h^{-1}(L) = \{x \varepsilon \Sigma^*: h(x) = y$ for some y in $L\}$.

(v) If $\tau = s_1, s_2, s_3, \ldots$ is a sequence of objects and $i_1, i_2, i_3, \ldots$ are such that $i_1 < i_2 < i_3 < \ldots$, then $s_{i_1}, s_{i_2}, s_{i_3}, \ldots$ is called a subsequence of τ.

(vi) $\emptyset$ denotes the empty set and Λ denotes the empty word.

(vii) If A is a (nondeterministic) finite automaton, then $L(A)$ denotes its language.

(viii) If A is an ultimately periodic sequence (set) of nonnegative integers then thres(A) denotes the smallest integer j for which there exists a positive integer q such that, for every $i \geq j$, if i is in A then $(i+q)$ is in A. The smallest positive integer p such that, for every $i \geq$ thres(A), whenever i is in A then also $(i+p)$ is in A, is denoted by per(A).

2. DEFINITIONS AND EXAMPLES OF EOL SYSTEMS

In this section we give basic definitions concerning developmental languages, which are relevant for this paper.

Definition 1. An EOL system is a construct $G = \langle V_N, V_T, P, \omega\rangle$ such that
V_N is a finite alphabet (of nonterminal letters and symbols),
V_T is a finite nonempty alphabet (of terminal letters or symbols), such that $V_N \cap V_T = \emptyset$,
ω is an element of $(V_N \cup V_T)^+$ (called the axiom of G),
P is a finite nonempty set (called the set of productions of G) each element of which is of the form $a \to \alpha$, where the symbol "$\to$" is not in $V_N \cup V_T$, a is in $V_N \cup V_T$, and α is in $(V_N \cup V_T)^*$. Moreover, for every a in $V_N \cup V_T$ there exists a word α in $(V_N \cup V_T)^*$ such that $a \to \alpha$ is in P.

In the sequel we shall often write "$a \underset{P}{\to} \alpha$" rather than "$a \to \alpha$ is in P." Also a production of the form $a \to \alpha$ is called a production for a in P.

Definition 2. An EOL system $G = \langle V_N, V_T, P, \omega\rangle$ is called a OL system if, and only if, $V_N = \emptyset$. (In this case we write G as $\langle V_T, P, \omega\rangle$).

OL systems are investigated, for example, in [Rozenberg and Doucet].

Definition 3. Let $G = \langle V_N, V_T, P, \omega\rangle$ be an EOL system.

(i) Let $x \in (V_N \cup V_T)^+$, say $x = b_1 \dots b_t$ for some $b_1, \dots, b_t$ in $V_N \cup V_T$, and let $y \in (V_N \cup V_T)^*$. We say that x directly derives y (in G), denoted as $x \underset{G}{\Longrightarrow} y$, if there exists a sequence $\pi_1, \dots, \pi_t$ of productions from P, such that, for every i in $\{1, \dots, t\}$, $\pi_i = b_i \to \alpha_i$ and $y = \alpha_1 \dots \alpha_t$.

(ii) As usual, $\overset{+}{\underset{G}{\Longrightarrow}}$ denotes the transitive closure of the relation $\underset{G}{\Longrightarrow}$ and $\overset{*}{\underset{G}{\Longrightarrow}}$ denotes the reflexive and transitive closure of the relation $\underset{G}{\Longrightarrow}$. If $x \overset{*}{\underset{G}{\Longrightarrow}} y$ then we say that x *derives* y *in* G.

(iii) A finite sequence $D = (x_0, x_1, \ldots, x_r)$ of words from $(V_N \cup V_T)^*$ such that, $r \geq 1$ and, for each i in $\{1, \ldots, r\}$, $x_{i-1} \underset{G}{\Longrightarrow} x_i$, is called a *derivation* (*of* x_r *from* x_0) *in* G. If $x_0 = \omega$, then D is called a *derivation of* x_r *in* G.

(iv) An infinite sequence $D = (x_0, x_1, \ldots)$ of words from $(V_N \cup V_T)^+$ such that, for each $i \geq 1$, $x_{i-1} \underset{G}{\Longrightarrow} x_i$, is called an *infinite derivation* in G.

(v) If $D = (x_0, x_1, \ldots, x_r)$ is a derivation in G, then its *control sequence* is any sequence $\tau = (T_1, \ldots, T_r)$ of subsets of P, such that, for each i in $\{1, \ldots, r\}$, $x_{i-1} \underset{G}{\Longrightarrow} x_i$ "using" all and only productions from T_i.

(vi) For x in $(V_N \cup V_T)^+$, y in $(V_N \cup V_T)^*$ and a positive integer r we write $x \overset{r}{\underset{G}{\Longrightarrow}} y$ if there exists a derivation $D = (x_0 = x, x_1, \ldots, x_r = y)$ in G. We also write $x \overset{0}{\underset{G}{\Longrightarrow}} x$, for every x in $(V_N \cup V_T)^+$.

(vii) max(G) is defined as $\max\{|\alpha| : a \underset{P}{\Longrightarrow} \alpha$ for some a in $V_N \cup V_T$ and α in $(V_N \cup V_T)^*\}$.

Definition 4. Let $G = \langle V_N, V_T, P, \omega \rangle$ be an EOL system. The *language of* G, denoted as L(G), is defined by $L(G) = \{x \in V_T^* : \omega \overset{*}{\underset{G}{\Longrightarrow}} x\}$.

Definition 5. A language K is called an *EOL language* (*OL language*) if, and only if, there exists an *EOL system* (*OL system*) G such that L(G) = K.

Remark 1. Given an EOL system $G = \langle V_N, V_T, P, \omega \rangle$, a derivation $D = (x_0, \ldots, x_r)$ and its control sequence $\tau = (T_1, \ldots, T_r)$, the pair (D, τ), in general, does not tell us which productions are used to rewrite the particular occurrences of letters in the words

$x_0,\ldots,x_{r-1}$. However (to avoid cumbersome notation and to keep the size of this paper decent) we shall often assume that the pair (D, τ) provides such information. This should not lead to confusion.

Remark 2. The properties we are interested in (in this paper) are trivial for finite languages, hence we consider only infinite languages. Thus in the sequel if we write "a language" (or "an EOL language") we mean an infinite one, unless explicitly stated otherwise. Also whenever we write "an EOL system" we mean one generating an infinite language.

Remark 3. Given an EOL system $G = \langle V_N, V_T, P, \omega\rangle$ we shall sometimes consider P to be the "set of names for productions" rather than the set of productions itself. In this sense we can talk about the words over P, etc., and this should not lead to confusion.

We end this section with two examples of EOL systems.

Example 1. $G = \langle V_N, V_T, P, \omega\rangle$, where $V_N = \{S\}$, $V_T = \{a, b\}$, $P = \{S \to a, S \to b, a \to a^2, b \to b^3\}$ and $\omega = S$, is an EOL system such that $L(G) = \{a^{2^n} : N \geq 0\} \cup \{b^{3^n} : n \geq 0\}$.

Example 2. $G = \langle \Sigma, P, \omega\rangle$, where $\Sigma = \{a, b\}$, $P = \{a \to (ab)^2, b \to \Lambda\}$ and $\omega = ab$, is a OL system such that $L(G) = \{(ab)^{2^n} : n \geq 0\}$.

3. BASIC NOTIONS AND THEIR PROPERTIES.

In this section we introduce basic notions describing the structure of OL languages we are interested in and we prove some properties of these notions.

Definition 6. Let L be a language over an alphabet Σ and let B be a nonempty subset of Σ. Let $I_{L,B} = \{n \in N :$ there exists a word w in L such that $\#_B(w) = n\}$.

(i) B is numerically dispersed (in L) if, and only if, $I_{L,B}$ is infinite and for every positive integer k there exists a positive integer n_k such that, for every u_1, u_2 in $I_{L,B}$, if $u_1 \neq u_2$, $u_1 > n_k$ and $u_2 > n_k$ then $\mathrm{abs}(u_1-u_2) > k$.

(ii) B is clustered (in L) if, and only if, $I_{L,B}$ is infinite and there exist positive integers k_1, k_2 such that $k_1 > 1$, $k_2 > 1$ and, for every word w in L, if $\#_B(w) \geq k_1$, then w contains at least two occurrences of symbols from B which are of distance smaller than k_2.

Definition 7. Let L be a language over an alphabet Σ and let a be in Σ. The symbol a is said to be frequent (in L) if, and only if, for every positive integer n there exists a word w in L such that $\#_a(w) > n$; otherwise a is called nonfrequent (in L).

Definition 8. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system, let B be a nonempty subset of Σ and let a be in Σ.

(i) We define a B-characteristic sequence of a (in G), denoted as Seq(G, B, a), as an infinite sequence $Z_1, Z_2, \ldots$ of finite subsets of N such that, for each $i \geq 1$ and every nonnegative integer n, n is in Z_i if, and only if, $a \overset{i}{\underset{G}{\Longrightarrow}} w$ for some w in Σ^* such that $\#_B(w) = n$.

(ii) $\mathrm{Seq}(G, B, a) = Z_1, Z_2, \ldots$ is called unique if, and only if, for every $i \geq 1$, $\#Z_i = 1$.

(iii) $\mathrm{Seq}(G, B, a) = Z_1, Z_2, \ldots$ is called bounded if, and only if, there exists a constant C such that, for every $i \geq 1$, $n < C$ for every n in Z_i. In this case we also say that a is B-bounded (in G) and that C bounds $\mathrm{Seq}(G, B, a)$. We say that a is B-unbounded (in G) otherwise.

(iv) $\mathrm{Seq}(G, B, a) = Z_1, Z_2, \ldots$ is called constant if, and only if, for each $i, j \geq 1$, $Z_i = Z_j$. In this case we also say that a is B-constant (in G).

Lemma 1. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system, let B be a nonempty subset of Σ and let a be a symbol in Σ. Let $\mathrm{Seq}(G, B, a) = Z_1, Z_2, \ldots$ and let $U(G, B, a) = \{i_1, i_2, \ldots\}$ be the set of positive integers such that, for every $j \geq 1$, j is in $U(G, B, a)$ if, and only if, $Z_j \neq 0$. Then $U(G, B, a)$ is an ultimately periodic set.

Proof.

Let G, B, a, $\mathrm{Seq}(G, B, a)$ and $U(G, B, a)$ be as in the statement of the lemma.

Let $A = \langle Q, V, \delta, q_0, F \rangle$ be a finite automaton such that

$Q = \Sigma$,

$V = P$,

$q_0 = a$,

$F = B$,

for every q, $\bar{q}$ in Q and every v in V, $\bar{q} \in \delta(q, v)$ if, and only if, v is a production of the form $q \to \gamma_1 \bar{q} \gamma_2$ for some γ_1, γ_2 in Σ^*.

We leave to the reader the easy proof of the fact that, for every $j \geq 1$, $Z_j \neq \{0\}$ if, and only if, there exist a word y over V such that $|y| = j$ and $\delta(q_0, y) \cap F \neq \emptyset$. Hence, $U(G, B, a) = \{n \in N :$ there exists a word y in L(A) such that $|y| = n\}$.

But it is well known (see, e.g., [Ginsburg, Theorems 2.1.2 and 2.1.3]) that the set of lengths of a regular language is an ultimately periodic set and consequently Lemma 1 holds.

Definition 9. Let $G = \langle \Sigma, P, \omega \rangle$ and let B be a nonempty subset of Σ. A B-uniform period of G, denoted as m(G, B) is defined to be the smallest positive integer such that

(i) for every b in Σ, $m(G, B) > \text{thres}(U(G, B, b))$, and

(ii) m(G, B) is divisible by per(G, B, b), for every b in Σ, such that U(G, B, b) is infinite.

Lemma 2. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and let B be a nonempty subset of Σ. If B is numerically dispersed in L(G), then, for every symbol a which is frequent in L(G), Seq(G, B, a) is unique.

Proof.

Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and let B be a numerically dispersed (in L(G)) subset of Σ. Let a be a symbol from Σ which is frequent in L(G).

Let us assume, to the contrary, that $\text{Seq}(G, B, a) = Z_1, Z_2, \ldots$ is not unique, meaning that, for some $i_0 \geq 1$, Z_{i_0} contains at least two nonnegative integers n_1 and n_2 (say $n_2 > n_1$, so that $n_2 > 0$).

Now, let m be an arbitrary positive integer.

Let x be a word in L(G) which contains t occurrences of the letter a for some $t \geq m_0$ (recall that a is frequent in L(G) and so such a word x exists).

Let $D = (x_0 = x, x_1, \ldots, x_{i_0})$ be a derivation of some word x_{i_0} from the word x with a control sequence $\tau = T_1, \ldots, T_{i_0}$ which is such that each occurrence of a in x contributes n_2 occurrences of symbols from B in x_{i_0} (recall that $n_2 \varepsilon Z_{i_0}$).

Let $\overline{D} = (x_0 = x, \overline{x}_1, \ldots, \overline{x}_{i_0})$ be a derivation of some word $\overline{x}_{i_0}$ from the word x with a control sequence $\overline{\tau} = \overline{T}_1, \ldots, \overline{T}_{i_0}$ which is such that exactly one occurrence of a in x contributes n_1 occurrences of symbols from B in $\overline{x}_{i_0}$ (recall that $n_1 \varepsilon Z_{i_0}$) and all other occurrences in x of the letter a as well as all occurrences in x of all other letters "behave" in exactly the same way as in the derivation D with the control sequence τ.

Thus $\#_B(x_{i_0}) - \#_B(\overline{x}_{i_0}) = n_2 - n_1$, where $\#_B(x_{i_0}) \geq t \cdot n_2$.

Consequently, there exists a positive integer k_0 (put $k_0 = n_2 - n_1 + 1$) such that for every integer m there exist integers u_1 and u_2 larger than m (put $u_1 = tn_2 - (n_2 - n_1)$, $u_2 = t \cdot n_2$ and notice that u_2 and u_1 are large enough if t is large enough) such that $u_2 > u_1$ (note that $u_2 - u_1 = n_2 - n_1$), u_1 and u_2 are in $I_{L(G),B}$ and $u_2 - u_1 < k_0$. Thus B is not numerically dispersed; a contradiction.

Hence Seq(G, B, a) must be unique and Lemma 2 holds.

Lemma 3. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and let B be a subset of Σ which is numerically dispersed in L(G). Let a be a symbol from Σ such that a is frequent in L(G) and a is B-unbounded in L(G). Then, for every α in Σ^*, if $a \xrightarrow[P]{} \alpha$, then α must contain at least one occurrence of a B-unbounded letter.

Proof.

Let G, B, and a satisfy the statement of the lemma.

Let Seq(G, B, a) = $Z_1, Z_2, \ldots$.

Let $a \xrightarrow[P]{} \alpha$ and let us assume, to the contrary, that either $\alpha = \Lambda$ or $\alpha \neq \Lambda$ and each letter which occurs in α is B-bounded.

First, let us note that, if for every γ such that $a \xrightarrow[P]{} \gamma$ we would have that either $\gamma = \Lambda$ or $\gamma \neq \Lambda$ and every letter which occurs in γ is B-bounded, then a itself would be B-bounded; a contradiction.

Thus for some $\overline{\alpha}$ in Σ^+ we have that $a \xrightarrow[P]{} \overline{\alpha}$ and $\overline{\alpha}$ contains an occurrence of a B-unbounded letter.

Let $\overline{G} = \langle \Sigma, \overline{P}, \omega \rangle$ be a OL system such that $\overline{P}$ differs from P only in this that all productions for a that are different from $a \to \alpha$ are delted. Then, obviously, $\text{Seq}(\overline{G}, B, a) = \overline{Z}_1, \overline{Z}_2, \ldots$ is bounded (say for every $i \geq 1$, if $n \in \overline{Z}_i$ then $n < \overline{C}$ for some positive integer constant $\overline{C}$ dependent on $\overline{G}$ only).

As a is B-unbounded, for every positive integer C (in particular for $\overline{C}$), there exists an integer i_C such that Z_{i_C} contains an integer larger than C. But, obviously, for each $i \geq 1$, $\overline{Z}_i \subseteq Z_i$, and consequently, for some $i_0 \geq 1$ (set $i_0 = i_{\overline{C}}$), Z_{i_0} must contain at least two different integers (one smaller than $\overline{C}$ and another equal to or larger than $\overline{C}$).

Consequently, Seq(G, B, a) is not unique which contradicts Lemma 2.

Thus, for every α in Σ^*, if $a \xrightarrow[P]{} \alpha$, then α must contain an occurrence of a B-unbounded letter. Hence Lemma 3 holds.

Definition 10. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and n a positive integer such that $n \geq 2$. Let $A(G, n) = \{x \in L(G)\text{: for some } i \text{ in } \{0, \ldots, n-1\},\ \omega \overset{i}{\underset{G}{\Longrightarrow}} x\}$.

(i) The n-decomposition of G, denoted as Dec(G, n), is the set of all OL systems of the form $\langle \Sigma, P^{(n)}, z \rangle$, where $z \in A(G, n)$ and $P^{(n)} = \{a \to \alpha : a \overset{n}{\underset{G}{\Longrightarrow}} \alpha\}$. Each OL system from Dec(G, n) is called a n-component of G.

(ii) A set $\{G_1, \ldots, G_p\}$ of OL systems is called a decomposition of G if, and only if, for some $n \geq 2$, $\{G_1, \ldots, G_p\} = \text{Dec}(G, n)$.

We leave to the reader the obvious proof of the following result.

Lemma 4. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and let $\mathcal{G} = \{G_1, \ldots, G_p\}$ be its decomposition. Let n be a positive integer such that $n \geq 2$ and let, for i in $\{1, \ldots, p\}$, $\mathcal{G}^{(i)} = \{G_1^{(i)}, \ldots, G_{p_i}^{(i)}\}$ be the n-decomposition of G_i. Then $\bigcup_{i=1}^{p} \mathcal{G}^{(i)}$ is a decomposition of G.

The following lemma states a number of properties that follow directly from defintions 6, 7, 8 and 10. As the proof of these properties is straightforward, we leave it to the reader. Some of these properties are so obvious and useful that they will be used in the sequel without directly quoting them.

Lemma 5. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and $\{G_1, \ldots, G_p\}$ be a decomposition of G.

(i) $L(G) = \bigcup_{i=1}^{p} L(G_i)$.

(ii) If B is a subset of Σ such that B is numerically dispersed in L(G), then, for every i in $\{1, \ldots, p\}$, if $I_{L(G_i),B}$ is infinite, then B is numerically dispersed in $L(G_i)$, and for at least one j in $\{1, \ldots, p\}$, B is numerically dispersed in $L(G_j)$.

(iii) If B is a subset of Σ such that, for every i in $\{1, \ldots, p\}$, B is either clustered in $L(G_i)$ or $I_{L(G_i),B}$ is finite, then B is clustered in L(G).

(iv) If $\{G_1, \ldots, G_p\}$ is a n-decomposition of G, B is a nonempty subset of Σ and a is in Σ, then, for each i in $\{1, \ldots, p\}$, $\mathrm{Seq}(G_i, B, a) = \overline{Z}_1, \overline{Z}_2, \ldots$ with $\overline{Z}_j = Z_{j \cdot n}$ for every $j \geq 1$ (where $\mathrm{Seq}(G, B, a) = Z_1, Z_2, \ldots$).

(v) If B is a nonempty subset of Σ and the letter a is B-bounded in G, then, for every i in $\{1, \ldots, p\}$, a is also B-bounded in G_i.

(vi) If the letter a is nonfrequent in L(G), then, for each i in $\{1,\ldots,p\}$, a is also nonfrequent in $L(G_i)$.

(vii) If the letter a is frequent in $L(G_i)$ for some i in $\{1,\ldots,p\}$, then a is also frequent in $L(G_i)$.

(viii) If B is a nonempty subset of Σ and a is a symbol from Σ such that a is B-unique (B-constant) in G, then, for each i in $\{1,\ldots,p\}$, a is B-unique (B-constant) in G_i.

Definition 11. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and $\mathcal{G} = \{G_1,\ldots,G_p\}$ be a decompositon of G. Let B be a nonempty subset of Σ. $\mathcal{G}$ is called a B-fitted decomposition of G if, and only if, for every i in $\{1,\ldots, p\}$ and for every a in Σ, if a is frequent in $L(G_i)$, then

(i) If a is B-bounded in G_i, then a is B-constant in G_i, and

(ii) If a is B-unbounded in G_i, then, for every $j \geq 1$, $Z_j = \{z_j\}$ for some $z_j \geq 2$ (where $\mathrm{Seq}(G_i, B, a) = Z_1, Z_2,\ldots$).

The reader may easily notice that, Lemma 2 implies that, if G is a OL system, B is numerically dispersed in L(G), $\mathcal{G}$ is a B-fitted decomposition of G and H is in $\mathcal{G}$, then if a is a symbol which is both, frequent in L(H) and B-bounded in H, then $\mathrm{Seq}(H, B, a) = \{z\}, \{z\},\ldots$ where z is a nonnegative integer.

4. THE EXISTENCE OF B-FITTED DECOMPOSITIONS.

In this section we prove that, for every OL system G and for every B which is numerically dispersed in L(G), there exists a B-fitted decomposition of G.

Lemma 6. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and let B be a nonempty subset of Σ such that B is numerically dispersed in L(G). Let $\mathcal{G} = \{G_1,\ldots,G_p\}$ be a m(G, B)-decomposition of G. Let i be in $\{1,\ldots,p\}$ and let a be in Σ. If a is frequent in $L(G_i)$ and $Seq(G_i, B, a) = Z_1, Z_2,\ldots,$ then either $Z_j = \{0\}$ for every $j \geq 1$, or $Z_j = \{z_j\}$ where $z_j \neq 0$ for every $j \geq 1$.

Proof.

Let G, B, and $\mathcal{G} = \{G_1,\ldots,G_p\}$ be as in the statement of the lemma. Let i be in $\{1,\ldots,p\}$ and let a be a symbol from Σ such that a is frequent in $L(G_i)$.

By Lemma 2 and Lemma 5 (vii), Seq(G, B, a) is unique and hence, by Lemma 5 (viii), $Seq(G_i, B, a)$ is unique.

By Lemma 1, U(G, B, a) is ultimately periodic. Hence, by Lemma 5 (iv) and by the definition of m(G, B), for every i in $\{1,\ldots,g\}$, if $Seq(G_i, B, a) = Z_1, Z_2,\ldots,$ then, for every $j \geq 1$, $Z_j \neq \{0\}$ if, and only if, $Z_1 \neq \{0\}$.

Thus Lemma 6 holds.

Lemma 7. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and let B be a nonempty subset of Σ such that B is numerically dispersed in G. Let $\mathcal{G} = \{G_1,\ldots,G_p\}$ be the m(G, B)-decomposition of G. Let i be in $\{1,\ldots,p\}$ and let a be in Σ. If a is frequent in $L(G_i)$ and a is B-bounded in G_i, then $Seq(G_i, B, a)$ is an ultimately periodic sequence.

Proof.

Let G, B and $\mathcal{G} = \{G_1,\ldots,G_p\}$ satisfy the statement of the lemma. Let i be in $\{1,\ldots,p\}$ and let a be a symbol from Σ such that a is frequent in L(G) and a is B-bounded in G_i. Let Seq(G_i, B, a) = $Z_1, Z_2, \ldots$.

By Lemma 6, either Seq(G_i, B, a) = {0}, {0},... or Seq(G_i, B, a) consists of singletons different from {0} only.

If Seq(G_i, B, a) = {0}, {0}, {0},... then obviously Lemma 7 holds.

Thus let us assume that Seq(G_i, B, a) consists of singletons different from {0} only.

Let D = (a = $x_0, x_1, x_2, \ldots$) be an infinite derivation in G_i such that, for each $i \geq 1$, x_i contains at least one occurrence of a letter frequent in L(G_i) such that its B-characteristic sequence consists of singletons different from {0} only. (Obviously such a derivation exists.) Let, for $j \geq 1$, Y_j be a subset of all these letters from Min(x_j) that their B-characteristic sequences consists of singletons different from {0} only.

First, we shall prove that there exists a constant F such that, for every $j \geq 1$, $\#_{Y_j}(x_j) < F$. Let us put F to be a positive integer constant such that, for every $j \geq 1$ and every integer n, if n is in Z_j then $n < F$ (recall that a is B-bounded). Let us assume to the contrary, that, for some $j_0 \geq 1$, $\#_{Y_{j_0}}(x_{j_0}) \geq F$. Thus x_{j_0} has more than F occurrences of letters which are frequent in L(G_i) and B-characteristic sequences of which consist of singletons different from {0} only. (Recall Lemma 6 and the choice of Y_j for each $j \geq 1$). Consequently, each such letter contributes at least one occurrence of an element from B to x_{j_0+1} (and, in fact, to each next word in D), and so $\#_B(x_{j_0+1}) > F$ which contradicts the fact that Seq(G_i, B, a) is bounded by F.

Thus our claim holds.

Now, for $j \geq 1$, let $\bar{x}_j$ denotes the word resulting by erasing from x_j of all occurrences of all letters in Min(x_j) - Y_j. Note that among all words of the form $\bar{x}_j$, for $j \geq 1$, there is only a finite number of different words (because none of these words is longer than F). Hence for some j_1, j_2 such that $j_2 > j_1$ we have $\overline{x}_{j_2} = \overline{x}_{j_1}$. But note that, for each $j \geq 2$, $\overline{x}_j$ "is contained" in the contribution to x_j from x_{j-1}. This is so, because (obviously) an occurrence in x_j of a letter the B-characteristic sequence of which consists only of singletons different from {0} may be contributed from an occurrence in x_{j-1} of a letter the B-characteristic sequence of which consists of singletons different from {0} only.

Thus for some j_1, j_2 such that $j_2 > j_1$ we have $\overline{x}_{j_2} = \overline{x}_{j_1}$ and, for every g in $\{j_1, j_1 + 1,\ldots,j_2 - 1\}$, we have $\overline{x}_g = \overline{x}_{g + s (j_2 - j_1)}$ for every $s \geq 0$.

But the "contributions to Seq(G_i, B, a)" from the words in the derivation D, depend (obviously) on the words $\overline{x}_j$ (for $j \geq 1$) only, and by the above, these contributions form an ultimately periodic sequence of numbers different from 0. However, (recall Lemma 6) the sequence Seq(G_i, B, a) consists of singletons only and so it is itself an ultimately periodic sequence.

Thus the lemma is proved.

Lemma 8. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system, let B be a nonempty subset of Σ such that B is numerically dispersed in L(G) and let $\mathcal{G} = \{G_1,\ldots,G_p\}$ be a m(G, B)-decomposition of G. Let $i \in \{1,\ldots,p\}$ and let a be a symbol from Σ such that a is frequent in $L(G_i)$ and a is B-unbounded in G_i. Let Seq(G_i, B, a) = Z_1, Z_2,... . Then for every positive integer C there exists a positive integer i_C such that, for every $j \geq i_C$, $Z_j = \{z_j\}$ where $z_j > C$.

Proof.

Let G, B, $\mathcal{G} = \{G_1,\ldots,G_p\}$, G_i, a and Seq(G_i, B, a) satisfy the statement of the lemma.

As a is frequent in $L(G_i)$ and B-unbounded in G_i, $I_{L(G_i),B}$ is infinite. Then by Lemmata 2, 5 (ii), 5 (vii) and 5 (viii), Seq (G_i, B, a) is unique, (say, for each $\ell \geq 1$, $Z_\ell = \{z_\ell\}$).

Let C be an arbitrary positive integer and let i_C be the smallest positive integer g such that $z_g > C \cdot \max(G)$.

Let $D = (a, x_1, \ldots, x_{i_C})$ be a derivation in G_i of a word x_{i_C} such that $\#_B(x_{i_C}) = z_{i_C}$. Let τ be a control sequence of D. Now, each occurrence of a letter from B in x_{i_C} must be derived (in (D, τ)) from some occurrence in $x_{i_C - 1}$ of a letter whose B-characteristic sequence in G_i consists of singletons different from $\{0\}$ only. Consequently (recall the choice of i_C) x_{i_C-1} must have more than C occurrences of letters whose B-characteristic sequences in G_i consist of singletons different from $\{0\}$ only. Hence, one can "prolongate" the derivation D to an infinite derivation $\overline{D} = (a, x_1, \ldots, x_{i_C}, x_{i_C+1}, \ldots)$ such that, for each $g \geq i_C$, $\#_B(x_g) > C$.

But $\mathrm{Seq}(G_i, B, a)$ consists of singletons only, and so Lemma 8 holds.

The property stated in Lemma 8 carries over through decompositions of OL systems in the following way. (We leave to the reader the obvious proof of the next result.)

Lemma 9. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system, let B be a nonempty subset of Σ such that B is numerically dispersed in L(G) and let $\mathcal{G} = \{G_1, \ldots, G_p\}$ be a m(G, B)-decomposition of G. Let $i \in \{1, \ldots, p\}$, let a be a symbol from Σ such that a is frequent in $L(G_i)$ and a is B-unbounded in G_i. Let $\mathcal{G}^{(i)} = \{G_1^{(i)}, \ldots, G_{p_i}^{(i)}\}$ be a decomposition of G_i and let $j \in \{1, \ldots, p_i\}$. Let $\mathrm{Seq}(G_j^{(i)}, B, a) = Z_1, Z_2, \ldots$. Then for every positive integer C there exists a positive integer i_C such that, for every $g \geq i_C$, $Z_j = \{z_j\}$, where $z_j > C$.

Definition 12. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system, let B be a nonempty subset of Σ such that B is numerically dispersed in G and let $\mathcal{G} = \{G_1, \ldots, G_p\}$

be the m(G, B)-decomposition of G. We define

(i) $m_1(G, B)$ to be the smallest positive integer such that, for every i in $\{1,\ldots,p\}$ and for every a in Σ such that a is frequent in $L(G_i)$ and B-bounded in G_i, $m_1(G, B)$ is divisible by per(Seq(G_i, B, a)) and $m_1(G, B) >$ thres(Seq(G_i, B, a)).

(ii) $m_2(G, B)$ to be the smallest positive integer such that, for every i in $\{1,\ldots,p\}$ and for every a in Σ such that a is frequent in $L(G_i)$ and a is B-unbounded in G_i, $m_2(G, B) > i_2$ where i_2 is defined as in the statement of Lemma 8.

(iii) $n(G, B) = m_1(G, B)\cdot m_2(G, B)$.

Lemma 10. Let $G = \langle\Sigma, P, \omega\rangle$ be a OL system, let B be a nonempty subset of Σ such that B is numerically dispersed in G and let $\mathcal{G} = \{G_1,\ldots,G_p\}$ be the $(m(G, B)\cdot n(G, B))$-decomposition of G. Let $i \in \{1,\ldots,p\}$ and let a be a letter in Σ such that a is frequent in $L(G_i)$ and a is B-bounded in G_i. Then a is B-constant in $L(G_i)$.

Proof.

This result follows directly from Lemma 4, Lemma 5 (ii), Lemma 5 (vii), Lemma 5 (viii), Lemma 7 and the definition of n(G, B).

Lemma 11. Let $G = \langle\Sigma, P, \omega\rangle$ be a OL system and let B be a nonempty subset of Σ such that B is numerically dispersed in G. Then there exists a B-fitted decomposition of G.

Proof.

Let G and B satisfy the statement of the lemma. From Lemma 5, Lemma 7, Lemma 8, Lemma 9, Lemma 10, and the definition of n(G, B) it follows that the $(m(G, B)\cdot n(G, B))$-decomposition of G is a B-fitted decomposition of G.

Thus Lemma 11 holds.

5. THE MAIN RESULT AND ITS APPLICATIONS.

In this section we prove the main result of this paper which states that if L is an EOL language and B is numerically dispersed in L, then B is clustered in L. We also show some applications of this result.

<u>Definition 13</u>. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system and let B be a nonempty subset of Σ. A word x in L(G) is called <u>B-dispersed in</u> G if, and only if, $\#_B(x) \geq 2$ and every two occurrences in x of symbols from Σ are of distance larger than max(G).

<u>Lemma 12</u>. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system, let B be a nonempty subset of Σ such that B is numerically dispersed in G and let $\mathcal{G} = \{G_1, \ldots, G_p\}$ be a B-fitted decomposition of G. Let $i \in \{1, \ldots, p\}$ and let x be a word in L(G) such that x is B-dispersed in G_i. Then, if $D = (x_0 = \omega, x_1, \ldots, x_r = x)$ is a derivation of x in G_i, then, for every j in $\{0, \ldots, r-1\}$, x_j does not contain an occurrence of a letter which is frequent in $L(G_i)$ and B-unbounded in G_i.

<u>Proof</u>.

Let G, B, G_i and x satisfy the statement of the lemma.

Let us assume, to the contrary, that $D = (x_0 = \omega, x_1 ., , , . x_r = x)$ is a derivation of x in G_i, such that for some f in $\{0, \ldots, r-1\}$, x_f contains an occurrence of a letter, say c, which is frequent in $L(G_i)$ and B-unbounded in G_i. But then (recall a definition of a B-fitted decomposition), for every $j \geq 1$, $Z_j = \{z_j\}$ for some $z_j \geq 2$, where $Seq(G_i, B, a) = Z_1, Z_2, \ldots$.

Thus by Lemma 3, x_{r-1} contains an occurrence of a letter which is frequent in $L(G_i)$ and B-unbounded in G_i, and this occurrence will contribute at least two occurrences in x_r of letters from B. These two occurrences, obviously, must be of distance smaller than max(G) which contradicts the definition of x.

Thus Lemma 12 holds.

Lemma 13. Let $G = \langle \Sigma, P, \omega \rangle$ be a OL system, let B be a nonempty subset of Σ and let $\mathcal{G} = \{G_1, \ldots, G_p\}$, be a B-fitted decomposition of G. Let $i \in \{1, \ldots, p\}$. If for every positive integer n there exists a word y_n in $L(G_i)$ such that $\#_B(y_n) \geq n$ and y_n is B-dispersed in G_i, then B is not numerically dispersed in $L(G_i)$.

Proof.

Let G, B and $\mathcal{G} = \{G_1, \ldots, G_p\}$ satisfy the statement of the lemma. Let $i \in \{1, \ldots, p\}$ and let us assume that for every positive integer n there exists a word y_n in $L(G_i)$ such that $\#_B(y_n) \geq n$ and y_n is B-dispersed in G_i.

For each a in Σ such that a is nonfrequent in G_i, let bound (G_i, a) be the smallest positive integer larger than the maximal number of occurrences of a in any word in $L(G_i)$, and let $t(a) = \max\{z : z \in Z_2^{(a)}\}$ where $\mathrm{Seq}(G_i, B, a) = Z_1^{(a)}, Z_2^{(a)}, \ldots$. Let $F = \sum_{a \in N(G_i)} t(a) \cdot \mathrm{bound}(G_i, a)$, where $N(G_i)$ denotes the set of all letters from Σ which are nonfrequent in G_i.

Let n be larger than $(\max(G))^2 \cdot (\#_B(\omega))$ and let x be a word in $L(G_i)$ such that $\#_B(x) \geq n$ and x is B-dispersed in G_i.

Let $D = (x_0 = \omega, x_1, \ldots, x_r = x)$ be a derivation of x in G_i (note that by the choice of n, $r \geq 3$). Let $\beta_0 = \#_B(x_0)$, $\beta_1 = \#_B(x_1), \ldots, \beta_r = \#_B(x_r)$. (We assume also that we are given a control sequence τ of D). Let s be an integer in $\{2, \ldots, r\}$ such that $\beta_r \leq \beta_{r-1} \leq \cdots \leq \beta_s > \beta_{s-1}$ (such an s exists because of our choice of n).

By Lemma 12, x_{s-2} does not contain occurrences of letters which are frequent in $L(G_i)$ and B-unbounded in G_i. Consequently $\beta_{s-1} = U_1^{(s-1)} + U_2^{(s-1)}$ and $\beta_s = U_1^{(s)} + U_2^{(s)}$ where

$U_1^{(s-1)}$ is the number of such occurrences in x_{s-1} of letters from B which are contributed from occurrences in x_{s-2} of letters which are nonfrequent in $L(G_i)$,

$U_1^{(s)}$ is the number of such occurrences in x_s of letters from B which are contributed from occurrences in x_{s-2} of letters which are nonfrequent in $L(G_i)$,

$U_2^{(s-1)}$ is the number of such occurrences in x_{s-1} of letters from B which are contributed from occurrences in x_{s-2} of letters which are frequent in $L(G_i)$ and B-bounded in G_i, and

$U_2^{(s)}$ is the number of such occurrences in x_s of letters from B which are contributed from occurrences in x_{s-2} of such letters which are frequent in $L(G_i)$ and B-bounded in G_i.

Because $\mathcal{G}$ is B-fitted, $U_2^{(s-1)} = U_2^{(s)}$, and by the definition of F, $U_1^{(s-1)} < F$. Hence $\beta_s - \beta_{s-1} = U_1^{(s)} - U_1^{(s-1)} < F$.

Thus we have proved (recall that $\beta_s = \beta_r$) that for n "large enough" if x is a word in $L(G_i)$ such that $\#_B(x) \geq n$ and x is B-dispersed in G_i, then there exists a word $\bar{x}$ in $L(G_i)$ (set $\bar{x} = x_{s-1}$) such that $\#_B(x) - \#_B(\bar{x}) < F$ where F is a constant dependent on G_i only.

Consequently B is not numerically dispersed in G_i and Lemma 13 holds.

Lemma 14. Let K be a OL language over an alphabet Σ and let B be a nonempty subset of Σ. If B is numerically dispersed in K, then B is clustered in K.

Proof.

Let K and B satisfy the statement of the lemma. Let K = L(G) where $G = \langle \Sigma, P, \omega \rangle$ is a OL system. Let $\mathcal{G} = \{G_1, \ldots, G_p\}$ be a B-fitted decomposition of G (its existence is guaranteed by Lemma 11). By Lemma 5 (ii), for each i in $\{1,\ldots,p\}$, either B is numerically dispersed in $L(G_i)$ or $I_{L(G_i),B}$ is finite, and for every j in $\{1,\ldots,p\}$, if $I_{L(G_j),B}$ is infinite, then B is numerically dispersed in $L(G_j)$.

Let s be in $\{1,\ldots,p\}$ and let $I_{L(G_s),B}$ be infinite. So B is numerically dispersed in $L(G_s)$. If we assume that B is not clustered in G_s, then for every positive integer larger than 1 there exists a word y_n in $L(G_i)$ such that $\#_B(y_n) \geq n$ and y_n is B-dispersed in G_i. But then, by Lemma 13, B is not numerically dispersed in $L(G_s)$; a contradiction.

Thus, for every i in $\{1,\ldots,p\}$, either $I_{L(G_i),B}$ is finite or B is clustered in $L(G_i)$.

But then, by Lemma 5 (iii), B must also be clustered in L(G) and so if B is numerically dispersed in K then B is clustered in G, which proves the lemma.

The following result, which was proved in [Ehrenfeucht and Rozenberg] turns out to be a very useful one for this paper.

Lemma 15. For every EOL language L there exist a OL language K and a coding ψ such that $\psi(K) = L$.

We leave to the reader the easy proof of our next result.

Lemma 16. Let K and L be languages (over alphabets Σ and V respectively) and let ψ be a coding such that $\psi(K) = L$. Then

(i) If B is a subset of V such that B is numerically dispersed in L, then $\psi^{-1}(B)$ is numerically dispersed in K, and

(ii) If U is a subset of Σ such that U is clustered in K, then $\psi(U)$ is clustered in L.

Theorem. Let K be an EOL language over an alphabet Σ and let B be a nonempty subset of Σ. If B is numerically dispersed in K, then B is clustered in K.

Proof.

This result follows directly from Lemma 14, Lemma 15, and Lemma 16.

It should be clear to the reader that the Theorem may be used to prove that a considerable number of languages are not EOL languages. This, by itself, fills in a gap in the developmental systems theory (for a discussion, see [Herman] or [Herman and Rozenberg]).

We shall present now two examples of such application.

Example 3. (See [Herman]). The language $L = \{x \in \{a, b\}^* : \#_a(x) = 2^n$ for some $n \geq 0\}$ is not an EOL language. This is so because $\{a\}$ is numerically dispersed in L, but, at the same time, $\{a\}$ is not clustered in L.

Example 4. Let ψ be a function from N to N such that, for every n in N, $\psi(n) \geq n$. Let $L_\psi = \{y_0 a y_1 a \ldots y_{r-1}\ a y_r : r = 2^n$ for some $n \geq 0$ and, for each i in $\{0,\ldots,r\}$, $y_i \in \{b, c\}^*$ and $|y_i| = \psi(r)\}$.

Finally, let us discuss the "structural" character of the Theorem. The Theorem states the structural rather than numerical characterization of the subclass of EOL languages. This statement is not precise but it may be illustrated by the following. Whereas it was proved in Example 3 that the language $L = \{x \in \{a, b\}^* : \#_a(x) = 2^n$ for some $n \geq 0\}$ is not an EOL language, the language $\overline{L} = \{a^{2^n}\ b^m : n, m \geq 0\}$ is generated by the EOL grammar $\langle V_N, V_T, P, \omega\rangle$ such that $V_N = \{S\}$, $V_T = \{a, b\}$, $P = \{S \to a,\ S \to \Lambda,\ S \to sb,\ a \to a^2,\ b \to b\}$, $\omega = S$. But $\overline{L}$ results from L by an appropriate permutation of occurrences of letters in the words of L. Consequently, all "numerical characteristics" (such as the set of lengths, the number of occurrences of particular symbols, etc.) are the same for both languages and one of them is an EOL language, while the other is not an EOL language.

REFERENCES.

Ehrenfeucht, A., and Rozenberg, G., Codings of OL languages, submitted for publication.

Ginsburg, S., *The mathematical theory of context-free languages*, McGraw-Hill, 1966.

Herman, G., Closure properties of some families of languages associated with biological systems, *Information and Control*, to appear.

Herman, G., and Rozenberg, G., *Developmental systems and languages*, North-Holland Publ. Company, to appear.

Lindenmayer, A., Mathematical models for cellular interactions in development, Journal of Theoretical Biology, 1968, V. 18, 300-315.

Lindenmayer, A., and Rozenberg, G., Developmental systems and languages, *Proc. of the 4th ACM Symp. on Theory of Comp.*, 1972.

Rozenberg, G., and Doucet, P., On OL languages, *Information and Control*, 1971, V. 19, 302-318.

GR:cah