

A COMBINATORIAL PROPERTY OF
EOL LANGUAGES

by

A. Ehrenfeucht, G. Rozenberg, R. Verraedt

CU-CS-273-84

August, 1984

All correspondence to the second author.

A. Ehrenfeucht, Department of Computer Science, University of Colorado,
Boulder, CO 80309, USA.

G. Rozenberg, Institute of Applied Mathematics and Computer Science, Univer-
sity of Leiden, The Netherlands.

R. Verraedt, Technical University, Group T, Leuven, Belgium.

ANY OPINIONS, FINDINGS, AND CONCLUSIONS
OR RECOMMENDATIONS EXPRESSED IN THIS PUB-
LICATION ARE THOSE OF THE AUTHOR AND DO
NOT NECESSARILY REFLECT THE VIEWS OF THE
NATIONAL SCIENCE FOUNDATION.

Abstract.

Let Δ be an alphabet and Π its nontrivial binary partition. Each word over Δ can uniquely be decomposed in subwords (called blocks) consisting of letters of Π_i only, $i \in \{1,2\}$. Let $K \subseteq \Delta^*$. K has a long block property (with respect to Π), abbreviated as *LB*-property, if there exists a function $f : \mathbf{N}^+ \rightarrow \mathbf{N}^+$ such that for every $w \in K$ and every positive integer m the number of blocks of length at most m in w is bounded by $f(m)$. K has a clustered block property (with respect to Π), abbreviated as *CB*-property, if there exists a positive integer n_0 and a growing function $g : \mathbf{N}^+ \rightarrow \mathbf{N}^+$ such that for every $w \in K$ and for every positive integer m the blocks of length at most m can be covered by at most n_0 segments of length at most $g(m)$.

It is proved that a *CB*-property always implies a *LB*-property but not necessarily other way around. It is proved that an EOL language has a *LB*-property if and only if it has a *CB*-property.

Introduction.

A study of combinatorial properties of languages in various language classes constitutes an active and important research area within formal language theory. A typical result here is of the form: if K is a language of type X , then $P(K)$ where P is a combinatorial property of K . Such a property can be expressed directly, as e.g. in all kinds of pumping theorems, or indirectly (conditionally) getting then the following form: if K is a language of type X and $P_1(K)$, then $P_2(K)$ where P_1, P_2 are combinatorial properties. In the case of EOL languages properties of this kind can be found e.g. in [ER] and [EnR].

This paper is concerned with a combinatorial property of the letter-type concerning EOL languages. Let Δ be an alphabet and Π its nontrivial binary partition. Each word over Δ can uniquely be decomposed in subwords (called blocks) consisting of letters of Π_i only, $i \in \{1,2\}$. Let $K \subseteq \Delta^*$. K has a long block property (with respect to Π) if there exists a function $f: \mathbf{N}^+ \rightarrow \mathbf{N}^+$ such that for every $w \in K$ and every positive integer m the number of blocks of length at most m in w is bounded by $f(m)$. K has a clustered block property (with respect to Π) if there exist a positive integer n_0 and a growing function $g: \mathbf{N}^+ \rightarrow \mathbf{N}^+$ such that for every $w \in K$ and every positive integer m the blocks of length at most m can be covered by at most n_0 segments of length at most $g(m)$.

A CB-property always implies a LB-property but not necessarily the other way around. It is proved that an EOL language has a LB-property if and only if it has a CB-property. This result is proved by the "in depth" analysis of derivations in EOL systems and in this way we believe that this paper contributes to our understanding of the nature of derivations in EOL systems. Also we provide some applications of our main result. The first of this yields an example of a language which is an ETOL but not an EOL language. The second example is

given grammatically (using the grammatical mechanism of the so-called regular pattern grammars, see e.g. [KR1]). We prove this language to be not an EOL language, which allows one to prove an important strict inclusion in [KR2]; we can do this without knowing precisely the form of strings belonging to this language.

The paper is organized as follows.

In Section 1 we recall some basic terminology concerning EOL systems.

In Section 2 we introduce the long block and clustered block properties. We give some examples and prove that a clustered block property always implies a long block property but not the other way around. We also prove that:

if K_1, \dots, K_l are languages which have a clustered block property and $K \subseteq (K_1 \cup \dots \cup K_l)^{\leq k}$ where k is a positive integer then K has a clustered block property.....(*)

In Sections 3 through 5 we prove that if K is an EOL language which has a long block property then K has also a clustered block property. The proof - which turns out to be very technical - is organized as follows.

In Section 3 we derive a normal form for an EOL system G generating K . The useful letters of G are divided into different types: 1, 2, 3I and 3II such that $K \subseteq (M_0 \cup \bigcup_{\alpha} L(G_{\alpha}))^{\leq k}$ where M_0 is a finite language and α ranges over all useful letters of G . Moreover for all types except for type 3II it is proved that $L(G_{\alpha})$ has a clustered block property. In view of (*) it then suffices to prove that for a letter α of type 3II, $L(G_{\alpha})$ has a clustered block property. This is done in two steps.

In Section 4 it is proved that $L(G_{\alpha}) \subseteq (M_1 \cup M_2 \cup M_3 \cup M_4)^{\leq k}$ for a positive integer k where M_1 and M_2 are languages which have a clustered block property.

Finally in Section 5 it is proved that also M_3 and M_4 have a clustered block property which concludes the proof of the main result of the paper.

We end the paper by some applications of our main result. This is done in Section 6.

1. Preliminaries.

We assume the reader to be familiar with the basic theory of EOL systems and languages, e.g. in the scope of [RS]. In this section we recall some basic terminology concerning EOL systems, fixing in this way the notation for our paper. Also, some new notions are introduced.

For a finite set X , $\#X$ denotes the number of elements of X . \mathbf{N} denotes the set of nonnegative integers and \mathbf{N}^+ denotes the set of positive integers. For a finite subset X of \mathbf{N} , $\min X$ and $\max X$ denote the minimum and maximum of X respectively. An alphabet is a finite nonempty set of symbols. $f : \mathbf{N}^+ \rightarrow \mathbf{N}^+$ denotes a total function with domain \mathbf{N}^+ and range \mathbf{N}^+ ; f is called *growing* if for every $n \in \mathbf{N}^+$, $f(n) \geq n$.

Let Δ be an alphabet. Λ denotes the empty word. For a word $w \in \Delta^*$, $|w|$ denotes its length and $\text{alph } w$ denotes the set of letters occurring in w . For an alphabet Θ $\#_{\Theta} w$ denotes the number of occurrences of letters from Θ in w . For a nonnegative integer i , $w(i)$ denotes the i -th letter of w if $1 \leq i \leq |w|$ and $w(i) = \Lambda$ otherwise. If w is nonempty, then $\text{last } w$ denotes $w(|w|)$. Let $w = uv$ where $u, v \in \Delta^*$. Then u is a *prefix* of w and v is a *suffix* of w ; we write $u \text{ pref } w$ and $v \text{ suf } w$ respectively.

For a language K , $K^0 = \{\Lambda\}$ and for a nonnegative integer n , $K^{n+1} = K^n \cdot K$. For a nonnegative integer n , $K^{\leq n} = \bigcup_{i=0}^n K^i$.

An EOL system will be denoted as $G = (\Sigma, P, w, \Delta)$ where Σ is its *total alphabet*, Δ is its *terminal alphabet*, $w \in \Sigma^+$ is its *axiom*, and P is the *set of productions*. In the notation of an EOL system we use a production set instead of a finite substitution since this seems to be more plausible for the purpose of this paper.

If $\alpha \in \Sigma$ and $\alpha \xrightarrow{P} x$ belongs to P then $\alpha \xrightarrow{P} x$ is an α -*production* of G . The fact that $\alpha \xrightarrow{P} x$ belongs to P is often abbreviated as $\alpha \xrightarrow{P} x$.

If $\alpha \in \Sigma$ and G is as above then $G_\alpha = (\Sigma, P, \alpha, \Delta)$.

Since the problems concerned in this paper become trivial otherwise we consider infinite EOL systems only (i.e. EOL systems which generate infinite languages), unless explicitly clear otherwise.

Let $G = (\Sigma, P, w, \Delta)$ be an EOL system.

(1) A letter $\alpha \in \Sigma$ is called *recursive* if $\alpha \xrightarrow[G]{+} u\alpha v, uv \in \Sigma^*$. The set of recursive letters of G is denoted by $\text{rec } G$.

(2) G is called *propagating* if $\alpha \xrightarrow{P} x$ implies $x \neq \Lambda$. In this case we say that G is an EPOL system.

(3) G is called *synchronized* if for every $\alpha \in \Delta, \alpha \xrightarrow[G]{+} x$ implies $x \notin \Delta^*$.

Without loss of generality we can assume that if G is a synchronized EOL system then there exists a symbol $F \in \Sigma - \Delta$, the *synchronization symbol* of G , such that $F \rightarrow F$ is the only F -production of G and for each $\alpha \in \Delta, \alpha \rightarrow F$ is the only α -production of G . In the rest of this paper whenever we consider a synchronized EOL system it will be assumed that its synchronization symbol equals F .

(4) G is called *standard* if the following conditions hold:

- (i) $w = S \in \Sigma - \Delta$.
- (ii) G is propagating and synchronized;
- (iii) for each $\alpha \in \Sigma, \alpha \xrightarrow{P} x$ implies $S \notin \text{alph } x$;

If G is standard, then we use $us\ G$ to denote $\Sigma - (\Delta \cup \{S, F, \})$ and we call $us\ G$ the set of *useful symbols* of G .

Let $G = (\Sigma, P, w, \Delta)$ be an EOL system and let l be a positive integer. Let us recall that a *derivation in G (of length l leading from $x \in \Sigma^*$ to $y \in \Sigma^*$)* is a sequence $(x = x_0, x_1, \dots, x_l = y)$, such that $x_0 \xrightarrow[G]{+} x_1, x_1 \xrightarrow[G]{+} x_2, \dots, x_{l-1} \xrightarrow[G]{+} x_l$

together with a precise description of how all the occurrences in x_i are rewritten to obtain x_{i+1} for $0 \leq i \leq l-1$. Such a description can be formalized (see, e.g. [RS]). For the purpose of this paper it suffices to depict a derivation D by

$$D : x_0 \xrightarrow{G} x_1 \xrightarrow{G} x_2 \xrightarrow{G} \cdots \xrightarrow{G} x_l .$$

A derivation in G leading from w to $x \in \Delta^*$ is called a *successful derivation in G* .

To each derivation there corresponds a derivation tree; if a derivation tree of G corresponds to a successful derivation in G , then it is called a *successful derivation tree (in G)*.

If $\Sigma_1 \subseteq \Sigma$ and T is a derivation tree of G whose nodes are labelled by elements of Σ_1 , then T is called a *Σ_1 -labelled derivation tree of G* .

In addition to the rather standard notation and terminology concerning derivation trees we will also use the following.

For a tree T , $\text{height } T$ denotes its *height*.

For a node v of a derivation tree we will use $l(v)$ to denote the *label of v* .

Let G be a EPOL system and let T be a derivation tree in G of height l . Then for $0 \leq i \leq l$, $\text{set}_i T$ denotes the set of nodes whose distance to the root equals i , and $\text{result}_i T$ denotes the word which results from the sequence of all nodes (ordered from left to right) from $\text{set}_i T$ by replacing each node by its label. Whenever we omit the index i in the above notation, it is assumed that i equals $\text{height } T$.

2. Long Block and Clustered Block Property.

In this section we define two combinatorial properties of languages over $\{a,b\}^*$ forming the subject of investigation of this paper: a long block property and a clustered block property. We need a number of auxiliary notions first.

A segment of a word w is nothing else but an occurrence of a nonempty subword of w , i.e., a subword of w together with its position within w . Formally we have the following definition.

Definition 2.1. Let Δ be an alphabet and let $w \in \Delta^*$. A segment of w is a construct (u, k, l) where $u \in \Delta^+$, $k, l \in \mathbf{N}$, $1 \leq k \leq l \leq |w|$ and $u = w(k)w(k+1)\dots w(l)$. The set of segments of w is denoted by $SEG(w)$. ■

In the sequel the usual terminology concerning words will also be used for segments (e.g., the length of a segment (u, k, l) is defined as $|u|$); however, this should not lead to confusion.

Let x and y be two segments of w . We say that y covers x if x on its own is a segment of y . This definition is generalized to sets of segments as follows.

Definition 2.2. Let $X, Y \subseteq SEG(w)$. We say that X covers Y if for every $(u, k, l) \in Y$ there exists a segment $(u', k', l') \in X$ such that $k' \leq k$ and $l' \geq l$. ■

For a word over a two-letter alphabet $\{a,b\}$ we now introduce the concept of a block. Intuitively we call a block of $w \in \{a,b\}^*$ a segment of w consisting entirely of a 's or b 's, which cannot be extended. Formally we have the following definition.

Definition 2.3. Let $w \in \{a,b\}^*$. A block of w is a construct $(u, k, l) \in SEG(w)$ such that either $u \in a^+$ and $w(k-1), w(l+1) \neq a$ or $u \in b^+$ and $w(k-1), w(l+1) \neq b$.

The set of all blocks of w is denoted $BL(w)$. For a positive integer m we also denote $BL^m(w) = \{(u, k, l) \in BL(w) \mid |u| \leq m\}$. ■

We are now ready to introduce the definition of a long block property. Informally speaking, a language $K \subseteq \{a, b\}^*$ is said to have a long block property if there exists a function which bounds the number of blocks of length of at most m .

Definition 2.4. Let $K \subseteq \{a, b\}^*$. Then K has a *long block property*, written $K \in LB$ if there exists a function $f : \mathbb{N}^+ \rightarrow \mathbb{N}^+$ such that for every $w \in K$ and every positive integer m , $\#BL^m(w) \leq f(m)$. We also say that $K \in LB$ with *parameter* f . ■

Example 2.1. Let $K_1 = \{aba^2b^2a^3b^3 \dots a^n b^n \mid n \geq 1\}$,
 $K_2 = \{a^{i_1} b^{j_1} a^{i_2} b^{j_2} \dots a^{i_n} b^{j_n} \mid n \geq 1 \text{ and } i_l, j_l \geq l \text{ for } 1 \leq l \leq n\}$, and
 $K_3 = \{a^{i_1} b^{j_1} a^{i_2} b^{j_2} \dots a^{i_n} b^{j_n} \mid n \geq 1 \text{ and } i_l, j_l \geq 1 \text{ for } 1 \leq l \leq n\}$. $K_1, K_2 \in LB$ but $K_3 \notin LB$.

Proof. Clearly for every $w \in K_1$ and every positive integer m , $\#BL^m(w) \leq 2m$. Thus if we define $f : \mathbb{N}^+ \rightarrow \mathbb{N}^+$ by $f(m) = 2m$, then $K_1 \in LB$ with parameter f . One can easily see that $K_2 \in LB$ with the same parameter f . Since $(ab)^n \in K_3$ for every positive integer n , $\#BL^1(w)$ cannot be bounded by a function $f : \mathbb{N}^+ \rightarrow \mathbb{N}^+$. Hence $K_3 \notin LB$. ■

We now introduce another "block property": a clustered block property. Intuitively speaking, a language $K \subseteq \{a, b\}^*$ has a clustered block property if all blocks of length at most m of a word w are not arbitrary scattered but occur in "clusters". Formally we have the following definition.

Definition 2.5. Let $K \subseteq \{a, b\}^*$. Then K has a *clustered block property*, written $K \in CB$ if there exists a positive integer n_0 and a growing function

$g : \mathbf{N}^+ \rightarrow \mathbf{N}^+$ such that for every $w \in K$ and every positive integer m there exists a $X \subseteq \text{SEG}(w)$ such that the following conditions hold:

- (i) $\#X \leq n_0$;
- (ii) for every $z \in X$, $|z| \leq g(m)$;
- (iii) X covers $BL^m(w)$.

We also say that $K \in \text{CB}$ with parameters n_0 and g . ■

The above definition says that we can cover all blocks of length at most m of a word w by at most n_0 segments bounded in length by $g(m)$. We illustrate this definition by the following example.

Example 2.2. Let K_1 and K_2 be as in Example 2.1. Then $K_1 \in \text{CB}$ and $K_2 \notin \text{CB}$.

Proof. Let $w = aba^2b^2a^3b^3 \dots a^n b^n \in K_1$. Let m be a positive integer. Let $l = \min\{m, n\}$. Then the segment $aba^2b^2a^3b^3 \dots a^l b^l$ covers all elements of $BL^m(w)$. Its length equals $2(1+2+\dots+l) = l(l+1) \leq m(m+1)$. Consequently $K_1 \in \text{CB}$ with parameters 1 and $g(m) = m(m+1)$. The fact that $K_2 \notin \text{CB}$ is proved by contradiction as follows.

Assume that $K_2 \in \text{CB}$ with parameters n_0 and g . Consider $m = n_0 + 1$ and $w = ab^{g(n_0+1)+1} a^2 b^{g(n_0+1)+2} \dots a^{n_0+1} b^{g(n_0+1)+n_0+1}$. Obviously every element of BL^{n_0+1} consists of a 's only and $\#BL^{n_0+1}(w) = n_0 + 1$. Now let $X \subseteq \text{SEG}(w)$ satisfy conditions (i) through (iii) of Definition 2.5 for $m = n_0 + 1$ and w . Since for every $z \in X$, $|z| \leq g(n_0 + 1)$, every $z \in X$ can cover at most one element of $BL^m(w)$. Consequently $\#BL^{n_0+1}(w) \leq n_0$; a contradiction. Hence $K_2 \notin \text{CB}$. ■

As far as K_3 is concerned, we have $K_3 \notin \text{CB}$. This follows from the general property that each clustered block property implies a long block property as can be seen from the following theorem.

Theorem 2.1. Let $K \subseteq \{a, b\}^*$. Then $K \in CB$ implies $K \in LB$.

Proof. Assume that $K \in CB$ with parameters n_0 and g . Define $f : \mathbb{N}^+ \rightarrow \mathbb{N}^+$ by $f(m) = n_0 \cdot g(m)$. We claim that $K \in LB$ with parameter f . This is proved as follows. Let $w \in K$, let m be a positive integer and let $X \subseteq SEG(w)$ be such that it satisfies (i) through (iii) of Definition 2.5. Then the total length of all segments of X is not longer than $\#X$ times the maximal length of a segment in X , i.e. $n_0 \cdot g(m)$. Thus - because the length of a block is always positive - $\#BL^m(w) \leq n_0 \cdot g(m) = f(m)$ and consequently $K \in LB$ with parameter f . ■

Observe that from Examples 2.1 and 2.2 it follows that the converse of Theorem 2.1 does not hold.

In the following theorem we state the obvious fact that if for a language there exists a positive integer which limits the number of blocks in every word then this language has a long block property and a clustered block property.

The following result is obvious and hence given without a proof.

Theorem 2.2. Let $K \subseteq \{a, b\}^*$. If there exists a positive integer n such that for each $w \in K$, $\#BL(w) \leq n$, then $K \in CB$ (and hence $K \in LB$). ■

We proceed to investigate operations on languages which preserve the CB-property.

Lemma 2.1. Let $K_1, K_2 \subseteq \{a, b\}^*$. If $K_1, K_2 \in CB$ and $K \subseteq K_1 \cup K_2$, then $K \in CB$.

Proof. Let K_1, K_2 be as in the statement of the lemma. Let $K_1 \in CB$ with parameters n_1 and g_1 , and let $K_2 \in CB$ with parameters n_2 and g_2 . Then clearly $K \in CB$ with parameters $\max\{n_1, n_2\}$ and g where $g(n) = \max\{g_1(n), g_2(n)\}$ for a positive integer n . ■

Let X be a set of segments of a word w . Such a set is called disjoint if no two segments overlap or "touch" each other. From a set of segments which is not disjoint we can construct a disjoint set by combining into one segment those segments that either overlap or touch each other. Formally we have the following definition.

Definition 2.6. Let $X \subseteq \text{SEG}(w)$. X is *disjoint* if for every $(u_1, k_1, l_1), (u_2, k_2, l_2) \in X$ either $k_2 > l_1 + 1$ or $k_1 > l_2 + 1$.

The *join* of X , denoted $\text{JOIN}(X)$, is defined as follows.

- (i) $\text{JOIN}(X) \subseteq \text{SEG}(w)$, $\text{JOIN}(X)$, is disjoint and covers X .
- (ii) For every disjoint $Y \subseteq \text{SEG}(w)$ which covers X , Y also covers $\text{JOIN}(X)$.

Lemma 2.2. Let $K_1 \subseteq \{a, b\}^*$. Let k be a positive integer. If $K_1 \in \text{CB}$ and $K \subseteq K_1^k$, then $K \in \text{CB}$.

Proof. Let K_1, k be as in the statement of the lemma. Let $K_1 \in \text{CB}$ with parameters n_1 and g_1 . We will demonstrate now that $K \in \text{CB}$ with parameters $k \cdot n_1$ and g where, for all positive integers n , $g(n) = k \cdot n_1 \cdot g_1(n)$.

Let $w \in K$ and let m be a positive integer. Then $w = w_1 w_2 \cdots w_k$ where $w_i \in K_1$ for $1 \leq i \leq k$. Since $K_1 \in \text{CB}$ for each w_i , $1 \leq i \leq k$, there exists an $X_i \subseteq \text{SEG}(w_i)$ such that:

- (i) $\#X_i \leq n_1$;
- (ii) for every $z \in X_i$, $|z| \leq g_1(m)$; and
- (iii) X_i covers $BL^m(w_i)$.

For $1 \leq i \leq k$ let

$$X'_i = \left\{ (u, k + \sum_{j=1}^{i-1} |w_j|, l + \sum_{j=1}^{i-1} |w_j|) \mid (u, k, l) \in X_i \right\}.$$

Let $X = \text{JOIN}(\cup_{i=1}^k X'_i)$.

Then the following conditions hold:

$$(i) \#X \leq \#(\cup_{i=1}^k X_i) = \sum_{i=1}^k \#X_i \leq k \cdot n_1,$$

(ii) for every $z \in X$,

$$|z| \leq \#(\cup_{i=1}^k X_i) \cdot \max\{|w| \mid w \in \cup_{i=1}^m X_i\} \leq k \cdot n_1 \cdot g_1(m)$$

and

(iii) X covers $BL^m(w)$.

To see that (iii) holds consider a block $u \in BL^m(w)$. Thus either $u = u_j u_{j+1} \cdots u_{j+s}$, $1 \leq j \leq j+s \leq k$ where $u_j \text{ suf } w_j$, $u_{j+1} \text{ pref } w_{j+s}$ and $u_{j+l} = w_{j+l}$ for $j < l < j+s$, or u is a subword of u_j , $1 \leq j \leq k$.

Each of those u_i 's, $j \leq i \leq j+s$ is covered by a segment from X_i and so u is covered by a segment of $JOIN(\cup_{i=1}^k X_i)$. Thus $K \in CB$ with parameters $k \cdot n_1$ and g . ■

Lemma 2.1 and Lemma 2.2 yield the following theorem.

Theorem 2.3. Let k, l be positive integers. Let $K_1, \dots, K_l \subseteq \{a, b\}^*$ such that $K_1, \dots, K_l \in CB$. If $K \subseteq (K_1 \cup \cdots \cup K_l)^{\leq k}$, then $K \in CB$. ■

In the above we have restricted ourselves to languages over a two letter alphabet. This however is not a real restriction. Definitions 2.3 through 2.5 can be generalized to languages over an arbitrary alphabet Δ (with cardinality at least 2) and a binary partition π of Δ . In this way we get the following definitions.

Definition 2.7. Let Δ be an alphabet and let $\pi = (\Delta_1, \Delta_2)$ be a binary partition of Δ (i.e. $\Delta_1, \Delta_2 \neq \phi$, $\Delta_1 \cup \Delta_2 = \Delta$ and $\Delta_1 \cap \Delta_2 = \phi$).

Then a *block of w (with respect to π)* is a construct $(u, k, l) \in SEG(w)$ such that either $u \in \Delta_1^+$ and $w(k-1), w(l+1) \notin \Delta_1$ or $u \in \Delta_2^+$ and $w(k-1), w(l+1) \notin \Delta_2$.

The set of all blocks of w (with respect to π) is denoted $BL_\pi(w)$ ($BL(w)$ if π is

understood).

For a positive integer m we also use the notation

$$BL_{\pi}^m(w) = \{(u, k, l) \in BL_{\pi}(w) \mid |u| \leq m\}. \quad \blacksquare$$

Definition 2.8. Let $K \subseteq \Delta^*$ and let π be a binary partition of Δ . Then K has a *long block property* (with respect to π), written $K \in LB(\pi)$ or $K \in LB$ if π is understood if there exists a function $f : \mathbf{N}^+ \rightarrow \mathbf{N}^+$ such that for every $w \in K$ and every positive integer m , $\#BL_{\pi}^m(w) \leq f(m)$. We also say that $K \in LB(\pi)$ with parameter f . \blacksquare

Definition 2.9. Let $K \subseteq \Delta^*$ and let π be a binary partition of Δ . Then K has a *clustered block property* (with respect to π), written $K \in CB(\pi)$ or $K \in CB$ if π is understood, if there exists a positive integer n_0 and a growing function $g : \mathbf{N}^+ \rightarrow \mathbf{N}^+$ such that for every $w \in K$ and every positive integer m there exists a $X \subseteq SEG(w)$ such that the following conditions hold:

- (i) $\#X \leq n_0$;
- (ii) for every $z \in X$, $|z| \leq g(m)$;
- (iii) X covers $BL_{\pi}^m(w)$.

We also say that $K \in CB(\pi)$ with parameters n_0 and g . \blacksquare

The following theorem establishes the connection between arbitrary alphabets and two letter alphabets that concerns long and clustered block properties.

Theorem 2.4. Let $K \subseteq \Delta^*$ and let $\Pi = (\Delta_1, \Delta_2)$ be a binary partition of Δ . Let h be the homomorphism on Δ^* defined by $h(\alpha) = a$ for $\alpha \in \Delta_1$, and $h(\alpha) = b$ for $\alpha \in \Delta_2$, where a, b are two fixed different letters. Then

- (1) $K \in LB(\Pi)$ if and only if $h(K) \in LB(\{a\}, \{b\})$, and

(2) $K \in CB(II)$ if and only if $h(K) \in CB(\{a\}, \{b\})$.

Proof. (1) Obvious.

(2) If $K \in CB(II)$, then $h(K) \in CB(\{a\}, \{b\})$. This can be seen as follows. If n_0, g are parameters proving that $K \in CB(II)$, then the same n_0, g will prove that $h(K) \in CB(\{a\}, \{b\})$; for a word $h(w)$ we consider covering by $h(X)$.

If $h(K) \in CB(\{a\}, \{b\})$, then $K \in CB(II)$. This can be seen as follows. If n_0, g are parameters proving that $h(K) \in CB(\{a\}, \{b\})$ then the same n_0, g will prove that $K \in CB(II)$. Let $w \in K$ and let m be a positive integer. Consider $h(w)$ and $X \subset SEG(h(w))$ such that conditions (i) through (iii) from Definition 2.5 are satisfied for $h(w)$. Let $X' = \{(u, k, l) \in SEG(w) \mid (h(u), k, l) \in X\}$. Clearly X' satisfies conditions (i) through (iii) from Definition 2.5 for w . Since w and m were arbitrary, $K \in CB(II)$. ■

3. A Normal Form Using Basic Letter Types.

The aim of this paper is to prove that if $K \subseteq \{a,b\}^*$ is an EOL language and $K \in LB$, then $K \in CB$.

Let $G_0 = (\Sigma_0, P_0, S, \{a,b\})$ be an EOL system which generates K , where we assume that

G is a standard EOL system.....(1)

To prove that $K \in CB$ we proceed as follows. Observe that

$$K = L(G_0) \subseteq \left(\bigcup_{\alpha \in us G_0} L((G_0)_\alpha) \right) \cup M_0 \stackrel{\leq maxr G_0}{},$$

where M_0 is a finite language (consisting of all words which can be derived in one step from S). In view of Theorems 2.2 and 2.3 it suffices to prove that for each $\alpha \in us G_0$, $L((G_0)_\alpha) \in CB$. To this aim we first replace G_0 by an EOL system which satisfies some special properties. This is done in several steps ((a) through (e)).

(a) We start by replacing G_0 by an EOL system G_1 which meets condition (1) and in addition

G_1 is in binary normal form.....(1')

i.e., its productions have one of the following forms:

$N \rightarrow NN, N \rightarrow N, N \rightarrow t, t \rightarrow N$, where N stands for an arbitrary nonterminal and t stands for an arbitrary terminal (see, e.g., [MSW]).

(b) Next a nonempty word $w \in K$ is given a *type* depending on the number of blocks it contains. We have the following definition.

Definition 3.1. (i) A nonempty word $w \in K$ is

- of type 1 if $\#BL(w) = 1$,
- of type 2 if $\#BL(w) = 2$, and
- of type 3 if $\#BL(w) \geq 3$.

Note that type 1 words are words of the form either a^n or b^n , $n > 0$ (the former are referred to as type 1a and the latter as type 1b). Type 2 words are words of the form either $a^n b^m$ or $b^n a^m$, $n, m > 0$ (the former are referred to as type 2a and the latter as type 2b).

(ii) Let $G = (\Sigma, P, w, \{a, b\})$ be an EOL system generating K . We say that a letter $\alpha \in us G$ has type i , $i \in \{1a, 1b, 2a, 2b, 3\}$, if $\alpha \xrightarrow[G]{+} w$ and w is of type i .

We will now transform G_1 into an EOL system G_2 generating K where each symbol has at most one type. This is done as follows.

Let $\Sigma_2 = \{a, b\} \cup \{\alpha[i] \mid \alpha \in us G_1, i \in \{1a, 1b, 2a, 2b, 3\}\} \cup \{S, F\}$.

Let P_2 consists of the following productions.

(i) If $S \xrightarrow{P_1} AB$, $A, B \in us G_1$, then $S \xrightarrow{P_2} A[i]B[j]$ for all $i, j \in \{3, 2a, 2b, 1a, 1b\}$.

(ii) If $S \xrightarrow{P_1} A$, $A \in us G_1$, then $S \xrightarrow{P_2} A[i]$ for each $i \in \{3, 2a, 2b, 1a, 1b\}$.

(iii) If $A \xrightarrow{P_1} BC$, $A, B, C \in us G_1$, then

$A[3] \xrightarrow{P_2} B[3]C[i]$ for each $i \in \{3, 2a, 2b, 1a, 1b\}$,

$A[3] \xrightarrow{P_2} B[2a]C[i]$ for each $i \in \{3, 2a, 2b, 1a\}$,

$A[3] \xrightarrow{P_2} B[2b]C[i]$ for each $i \in \{3, 2a, 2b, 1b\}$,

$A[3] \xrightarrow{P_2} B[1a]C[i]$ for each $i \in \{3, 2b\}$,

$A[3] \xrightarrow{P_2} B[1b]C[i]$ for each $i \in \{3, 2a\}$,

$A[2a] \xrightarrow{P_2} B[2a]C[1b]$,

$A[2a] \xrightarrow{P_2} B[1a]C[i]$ for each $i \in \{2a, 1b\}$,

$A[2b] \xrightarrow{P_2} B[2b]C[1a]$,

$$A[2b] \xrightarrow{P_2} B[1b]C[i] \text{ for each } i \in \{2b, 1a\},$$

$$A[1a] \xrightarrow{P_2} B[1a]C[1a],$$

$$A[1b] \xrightarrow{P_2} B[1b]C[1b].$$

(iv) If $A \xrightarrow{P_1} B, A, B \in us G_1$, then $A[i] \xrightarrow{P_2} B[i]$ for each $i \in \{3, 2a, 2b, 1a, 1b\}$.

(v) If $S \xrightarrow{P_1} x, x \in \{a, b\}$ then $S \xrightarrow{P_2} x$.

(vi) If $A \xrightarrow{P_1} a, A \in us G_1$ then $A[1a] \xrightarrow{P_2} a$.

(vii) If $A \xrightarrow{P_1} b, A \in us G_1$ then $A[1b] \xrightarrow{P_2} b$.

(viii) $a \xrightarrow{P_2} F, b \xrightarrow{P_2} F$ and $F \xrightarrow{P_2} F$.

Clearly $L(G_2) = K, G_2$ inherits the properties (1) and (1') from G_1 and each $\alpha \in us G_2$ has at most one type.....(1'')

(c) Thirdly, following the lines of the standard proof of the theorem $L(EOL) = L(COL)$ (see, e.g., [RS]) we can transform G_2 into an EOL system G_3 which inherits properties (1) and (1'') and is such that

$$\text{for each } \alpha \in us G_3, S \xrightarrow[G_3]{*} u\alpha v \text{ for some } uv \in (us G_3)^*, \alpha \xrightarrow[G_3]{*} x \text{ for some } x \in (us G_3)^+ \text{ and } \alpha \xrightarrow[G_3]{*} y \text{ for some } y \in \{a, b\}^+ \dots\dots\dots(2)$$

Hence from (1'') and (4) it follows that

$$\text{each } \alpha \in us G_3 \text{ has exactly one type } \dots\dots\dots(3)$$

If an EOL system satisfies (3) we say that " α is of type i " rather than " α has type i ". For $x \in \{1a, 1b, 2a, 2b, 3\}$ $type x$ denotes the set $\{\alpha \mid type \alpha = \{x\}\}$. Furthermore, for a derivation tree, each node labelled by a letter of type x is called a *node of type x* .

(d) Let $\alpha \in us G_3 \cap (type 1 \cup type 2)$ and consider $\alpha \xrightarrow[G_3]{+} w \in (us G_3)^+$. Observe

that w contains at most one letter of type 2 and the rest of w are letters of type

1. Clearly:

either there exists a positive integer n_0 such that for every positive integer n

there exists a word $w_n \in (us G_3)^+$ such that $\alpha \xrightarrow[G_3]{n} w$ and $\#_{type 1a} w \leq n_0$

or such a positive integer n_0 does not exist.

An analogous situation is true for letters of type 1b.

Now using the standard speed-up technique (see, e.g., [RS]) one can construct an EOL system G_4 such that (1),(2) and (3) remain true and (see the above remark) furthermore

for each $\alpha \in us G_4 \cap (type 1 \cup type 2)$ either there exists a $n_0 > 0$ such that

for each positive integer n , $\alpha \xrightarrow[G_4]{n} w \in (us G_4)^+$ and $\#_{type 1a} w \leq n_0$ or for

every positive integer n , $\alpha \xrightarrow[G_4]{n} w \in (us G_4)^+$ implies $\#_{type 1a} w \geq n$; more-

over this property also holds if we replace 1a by 1b(4)

(e) Finally, again using the speed-up technique, it is possible to construct an EOL system $G = (\Sigma, P, S, \{a, b\})$ which satisfies (1) through (4) and in addition satisfies the following two properties.

If $\alpha \in us G \cap rec G$, then $\alpha \xrightarrow[G]{*} u\alpha v$ for some $uv \in (us G)^*$ (5)

Let $\alpha \in us G$ be of type 3 and such that $alph x$ does not contain recursive

symbols of type 3 whenever $\alpha \xrightarrow[G]{+} x$. Then $\alpha \xrightarrow[P]{+} y$ and $alph y \subseteq us G$

imply $alph y$ contains only letters of type 1 and type 2.....(6)

The above property (6) gives rise to the division of letters of type 3 into two categories.

Definition 3.2. The set type 3I is the set of all letters of type 3 such that

$\alpha \xrightarrow[G]{+} x$ implies $alph x \cap rec G \cap type 3 = \phi$ and type 3II is the set type 3 - type

3I.

Nodes of *type 3I* (*3II* respectively) are nodes labelled by letters of *type 3I* (*type 3II* respectively). ■

For the rest of the paper we fix an EOL system $G = (\Sigma, P, S, \{a, b\})$ which satisfies (1) through (6) and generates K .

Recall that K has a long block property and

$$K \subseteq \left(\left(\bigcup_{\alpha \in us G} L(G_\alpha) \right) \cup M_0 \right)^{\leq \max G}$$

where M_0 is a finite language.

We proceed now by proving that for each $\alpha \in us G$, $L(G_\alpha) \in CB$. The following theorem states the result for the case when $\alpha \in type 1 \cup type 2 \cup type 3I$.

Theorem 3.1. If $\alpha \in type 1 \cup type 2 \cup type 3I$, then $L(G_\alpha) \in CB$.

Proof. If α is either of *type 1* or of *type 2*, then each word in $L(G_\alpha)$ has at most two blocks. Thus by Theorem 2.2 $L(G_\alpha) \in CB$. If $\alpha \in type 3I$ then the result follows from the above and Theorem 2.3. ■

It remains to prove that for $\alpha \in type 3II$, $L(G_\alpha) \in CB$. This will be shown in the next two sections.

We end this section by proving an upper bound on the number of nodes of *type 3* on a level of a derivation tree.

Lemma 3.1. There exists a positive integer k such that, for every (*us G*)-labelled derivation tree T of G and every $0 \leq i \leq \text{height } T$, $\#_{type 3} \text{ result}_i T \leq k$.

Proof. Since $K \in LB$, there exists a function $f : \mathbb{N}^+ \rightarrow \mathbb{N}^+$ such that $\#BL^{\max G}(w) \leq f(\max G)$. Let $k = f(\max G)$. We prove that for this choice of k the lemma holds. This is proved by contradiction.

Assume that T is a ($us G$)-labelled tree of G and $0 \leq i \leq \text{height } T$ is such that $\#_{\text{type } 3} \text{result}_i T > k$. Clearly based on T a successful derivation tree T' of G can be constructed such that there exists a $0 \leq j \leq \text{height } T'$ such that $\#_{\text{type } 3} \text{result}_j T' > k$.

Then, since each letter of type 3 does create at least one block and since G satisfies (2),

$$S \xrightarrow[G]{j} \text{result}_j T' \xrightarrow[G]{} w \in K \text{ and } \#BL^{\text{maxr}G}(w) > k = f(\text{maxr}G);$$

a contradiction. Thus the result holds. \square

For the rest of the paper let k_1 be an arbitrary but fixed integer which satisfies the statement of Lemma 3.1.

4. Limbs and Spikes.

In the previous section we have divided letters of G into basic types and for each letter type except for *type* 3II we have proved that a letter of this type gives rise to a language in CB .

In this section we make a first step in proving that a letter of type 3II also gives rise to a language in CB . We will prove that there exist four languages, M_1 through M_4 , and a positive integer k , such that if $\alpha \in \text{type } 3II$, then $L(G_\alpha) \subseteq (M_1 \cup M_2 \cup M_3 \cup M_4)^{\leq k}$, where, moreover M_1 and $M_2 \in CB$. In the next section we prove that M_3 and M_4 also belong to CB . Then from Theorem 2.3 it follows that $L(G_\alpha) \in CB$.

Our division of $L(G_\alpha)$ into elements of M_1 through M_4 is based on a detailed analysis of successful derivation trees in G_α . To this aim we need quite a number of new definitions which will be introduced now.

Consider a successful derivation tree T in G_α and remove from it all nodes which are not of type 3II. We end up with a derivation tree as depicted in Figure 1. The remaining tree is called the *skeleton of T* and is denoted as *skel T* . Two kinds of nodes occur: nodes with at most one successor (in *skel T*) and nodes with more than one successor. The former are called *simple* denoted by \cdot in Figure 1); the latter are called *complex* (denoted by \circ in Figure 1).

The following lemma gives an upper bound for the number of complex nodes occurring in a successful derivation tree of G_α where $\alpha \in \text{type } 3II$.

Lemma 4.1. There exists a positive integer k such that for every $\alpha \in \text{type } 3II$ and every successful derivation tree T of G_α , *skel T* contains no more than k complex nodes.

Proof. The lemma is proved by contradiction. Assume that no bound k exists. Let $\alpha \in \text{type } 3II$ and let T be a successful derivation tree of G_α such that

skel T has more than k_1 complex nodes. Recall that k_1 is an arbitrary but fixed integer satisfying the statement of Lemma 3.1. This T can now easily be transformed (using property (5) of Section 3 and the definition of letters of type 3II) into a ($us G$)-labelled tree T' of G with at least k_1+1 complex nodes and such that each node of type 3II has a son of type 3II. Consequently there is a $i \geq 0$ such that $\#_{type\ 3} result_i T' > k_1$ which contradicts Lemma 3.1. Thus the lemma holds. ■

For the rest of the paper let k_2 be an arbitrary but fixed integer which satisfies the statement of Lemma 4.1.

We now proceed with the analysis of *skel T*. A maximal (looked at top-down) branch in *skel T* which contains only simple nodes and ends either on a complex node or on a leaf (of *skel T*) is called a limb (of T). In Figure 2 all limbs are indicated by boxes. Such a limb will be denoted as $\langle v_1, \dots, v_k \rangle$ where v_1, \dots, v_k are consecutive nodes of the limb. A *spike (of a limb)* is a subbranch of a limb consisting of nodes $v_l, \dots, v_m, m \geq l$ such that v_{m+1} still belongs to the same limb, and v_l and v_{m+1} have the same label (i.e. $l(v_l) = l(v_{m+1})$). A spike as above is said to be of *type* $l(v_l)$ and will be denoted by $\langle\langle v_l, \dots, v_m \rangle\rangle$.

The notion of a spike is illustrated in Figure 3. We have depicted a limb $\langle v_1, \dots, v_{15} \rangle$. To the left of it we have indicated the nodes and to the right of it we have indicated the labels.

- $\langle\langle v_1, v_2 \rangle\rangle, \langle\langle v_1, \dots, v_4 \rangle\rangle, \langle\langle v_1, \dots, v_8 \rangle\rangle, \langle\langle v_3, v_4 \rangle\rangle,$
 $\langle\langle v_3, \dots, v_8 \rangle\rangle$ and $\langle\langle v_5, \dots, v_8 \rangle\rangle$ are all spikes of type a .
- $\langle\langle v_4, \dots, v_{10} \rangle\rangle, \langle\langle v_4, \dots, v_{14} \rangle\rangle$ and $\langle\langle v_{11}, \dots, v_{13} \rangle\rangle$ are all spikes of type b .
- $\langle\langle v_2, \dots, v_9 \rangle\rangle, \langle\langle v_2, \dots, v_{13} \rangle\rangle$ and $\langle\langle v_{10}, \dots, v_{13} \rangle\rangle$ are all spikes of type c .
- $\langle\langle v_6, \dots, v_{11} \rangle\rangle$ is a spike of type d .
- $\langle\langle v_7, \dots, v_{12} \rangle\rangle$ is a spike of type e .
- Observe that there are no spikes of type f since the given limb contains only

one node with label f .

Next we consider nodes (and words) which are descendants of the nodes of a spike.

Definition 4.1 Let $\alpha \in \text{type III}$, let T be a successful derivation tree of G_α of height l , let $\rho = \langle v_{r_1}, \dots, v_{r_2} \rangle$ be a limb of T such that for $r_1 \leq i \leq r_2$, v_i has distance i to the root and let $\sigma = \langle\langle v_{i_1}, \dots, v_{i_2} \rangle\rangle$ be a spike of ρ .

Then for $i_1 < i \leq l$, $lset_i(T, \sigma)$ denotes the following subset of $set_i T$.

If $i_1 < i \leq i_2 + 1$ then $lset_i(T, \sigma)$ consists of all nodes of $set_i T$ which are descendants of v_{i_1}, \dots, v_{i-1} and which are to the left of v_i .

If $i_2 + 1 < i \leq l$ then $lset_i(T, \sigma)$ consists of all nodes of $set_i T$ which are descendants of v_{i_1}, \dots, v_{i_2} and which are to the left of all elements of $set_i T$ which are descendants of v_{i_2+1} .

If in the above definition we replace left by right then we get the definition of $rset_i(T, \sigma)$.

Finally $lresult_i(T, \sigma)$ ($rresult_i(T, \sigma)$ respectively) is the word which results from the sequence of all nodes (ordered from left to right) from $lset_i(T, \sigma)$ ($rset_i(T, \sigma)$ respectively) by replacing each node by its label.

If $i = \text{height } T$ then in the above the subscript i is omitted. ■

Figure 4 illustrates the above definition. We have depicted $T, \sigma, lresult_i(T, \sigma)$ and $rresult_i(T, \sigma)$ for various levels.

Based on the above definition we will define for a letter α of type III two languages, called the left contributions of α and the right contributions of α . Consider the situation as depicted in Figure 4. Moreover assume that v_{i_1} has label β . Then we say that x is a left contribution of β and y is a right contribution of β . Formally we have the following definition.

Definition 4.2. Let $\beta \in$ type 3II. We associate with β two languages $LC(\beta)$

- the *left contributions* of β , and $RC(\beta)$ - the *right contributions* of β as follows.

(1) $x \in LC(\beta)$

if and only if

there exists a letter α of type 3II, a successful derivation tree T of G_α , a limb $\rho = \langle v_1, \dots, v_k \rangle$ of T and a spike $\sigma = \langle\langle v_i, \dots, v_m \rangle\rangle$ of ρ of type β such that $x = lresult(T, \sigma)$.

(2) $x \in RC(\beta)$

if and only if

there exists a letter α of type 3II, a successful derivation tree T of G_α , a limb $\rho = \langle v_1, \dots, v_k \rangle$ of T and a spike $\sigma = \langle\langle v_i, \dots, v_m \rangle\rangle$ of ρ of type β such that $x = rresult(T, \sigma)$. ■

We now return to the situation depicted in Figure 3 - as we have seen, the given limb contains several spikes. Let us consider the contributions of the limb to the word generated - we have already a formalism to denote contributions (both left and right) of a spike. Based on it we define now unique splitting of a limb into (disjoint) spikes called true spikes.

Definition 4.3. Let $\alpha \in$ type 3II. Let T be a successful derivation tree of G_α and let $\rho = \langle v_1, \dots, v_k \rangle$ be a limb of T .

Let $SPIKE(\rho)$ be the (possibly empty) sequence of spikes defined inductively as follows.

If ρ does not have spikes then $SPIKE(\rho) = \emptyset$.

Otherwise $SPIKE(\rho) = (\rho_1, \dots, \rho_t)$, $t \geq 1$ where ρ_1, \dots, ρ_t are spikes constructed one-by-one as follows.

Choose the first (top-down) node v_{i_1} on ρ such that ρ contains another node labelled by the same letter ; let v_{j_1+1} be the last node on ρ labelled by the same

letter as v_{i_1} . Then $\rho_1 = \langle\langle v_{i_1}, v_{i_1+1}, \dots, v_{j_1} \rangle\rangle$. Then start with v_{j_1+1} and proceed as above. One stops when there are no spikes left in the remaining part of the limb.

By Rest ρ we denote the set of all nodes from ρ not involved in any of ρ_1, \dots, ρ_t . Elements ρ_1, \dots, ρ_t of $SPIKE(\rho)$ are called *true spikes of ρ* . For every limb ρ of T , a true spike of ρ is called a *true spike of T* . ■

Consider again Figure 3. The true spikes of the given limb ρ (indicated by boxes in Figure 5) are $\langle\langle v_1, \dots, v_8 \rangle\rangle$ and $\langle\langle v_{10}, \dots, v_{13} \rangle\rangle$; Rest ρ equals $\{v_9, v_{14}, v_{15}\}$. Observe that both, the number of true spikes of ρ and the cardinality of Rest ρ , are bounded by the number of letters in the alphabet of the given system.

In order to divide the result of a successful derivation tree starting from a letter α of type 3II into words of four different languages we need the following definition.

Definition 4.4. Let $\alpha \in$ type 3II and let T be a successful derivation tree of G_α . Let C_1 be the set of all nodes which occur in a true spike of T and let C_2 be the set of all nodes which contribute in one step to the terminal word. Observe that $C_1 \cap C_2 = \emptyset$ (the last node of a limb is never included in a spike). C_3 is the set of all nodes of *skel* T not contained in $C_1 \cup C_2$.

Then the *extended skeleton of T* , denoted *eskel* T is obtained from *skel* T by including also all nodes (and their edges to nodes from *skel* T) that can be obtained in one step from nodes in C_3 .

The nodes of *eskel* T which do not belong to *skel* T are referred to as *added nodes of T* . Observe that none of the added nodes is a leaf of T and all added nodes belong to type 1 \cup type 2 \cup type 3 I. ■

Figure 6 depicts a typical situation.

- \square are nodes of true spikes (C_1 nodes),
- \odot are C_2 nodes,
- \bullet are C_3 nodes,
- \circ are the nodes which belong to $eskel T$ but not to $skel T$ (the added edges are denoted by dotted lines).

$$\text{Let } M_1 = \bigcup_{\beta \in \text{type } 1 \cup \text{type } 2 \cup \text{type } 3I} L(G_\beta),$$

$$M_2 = \{x \in \{a, b\}^+ \mid \beta \xrightarrow{C} x \text{ and } \beta \in \text{type } 3II\},$$

$$M_3 = \bigcup_{\beta \in \text{type } 3II} LC(\beta), \text{ and}$$

$$M_4 = \bigcup_{\beta \in \text{type } 3II} RC(\beta).$$

Observe that $x = x_1 x_2 \cdots x_{11}$, where

$$x_1, x_2, x_4, x_7, x_{10} \in M_1;$$

$$x_5, x_8 \in M_2; x_3, x_6 \in M_3 \text{ and } x_7, x_{11} \in M_4.$$

Clearly the following lemma holds.

Lemma 4.2. Let $\alpha \in \text{type } 3II$ and let $M = M_1 \cup M_2 \cup M_3 \cup M_4$, where M_1, M_2, M_3 and M_4 are defined as above. Then $L(G_\alpha) \subseteq M^+$. \blacksquare

We can improve the result of Lemma 4.2 and prove the existence of a positive integer k such that $L(G_\alpha) \subseteq M^{\leq k}$. We need the following lemma which provides bounds on the number of true spikes of an arbitrary successful derivation tree of G_α and on the number of its added nodes ($\alpha \in \text{type } 3II$).

Lemma 4.3. Let $\alpha \in \text{type } 3II$ and let T be a successful derivation tree of G_α .

(1) The number of true spikes of T is bounded by $\# \Sigma \cdot (k_2 \cdot \max r G + 1)$.

(2) The number of added nodes of T is bounded by $\# \Sigma \cdot (k_2 \cdot \max r G + 1) \cdot$

$\max r G$.

Proof. (1) Since each limb of T either starts from a son of a complex node or from the root, the number of limbs of T is bounded by $k_2 \cdot \max r G + 1$. Since every limb can contain at most $\# \Sigma$ true spikes, the number of true spikes of T is bounded by $\# \Sigma \cdot (k_2 \cdot \max r G + 1)$.

(2) Clearly the number of added nodes of T is bounded by the number of limbs of T times the number of nodes on a limb not included in a true spike of T times $\max r G$, thus by $\# \Sigma \cdot (k_2 \cdot \max r G + 1) \cdot \max r G$. ■

We are now ready to prove the main result of this section.

Lemma 4.4. Let $\alpha \in \text{type III}$ and let M be as above. Then there exists a positive integer k such that $L(G_\alpha) \subseteq M^{\leq k}$.

Proof. Let $k = (k_2 \cdot \max r G + 1)(\# \Sigma \cdot (\max r G + 2) + 1)$. Consider a successful derivation tree T in G_α of w . The word w is divided into subwords as follows:

- (i) contributions of added nodes,
- (ii) contributions of nodes of the C_2 category,
- (iii) left contributions of true spikes of T ,
- (iv) right contributions of true spikes of T .

Obviously in this way $w = w_1 \cdots w_p$ where for $1 \leq i \leq p$, $w_i \in M$. To prove the lemma it suffices to prove that $p \leq k$.

(1) $\#\{i \mid 1 \leq i \leq p, w_i \in M_1\}$ is bounded by the number of added nodes of T thus by $\# \Sigma \cdot (k_2 \cdot \max r G + 1) \cdot \max r G$ (see Lemma 4.2 (2)).

(2) $\#\{i \mid 1 \leq i \leq p, w_i \in M_2\}$ is bounded by the number of nodes of the C_2 category, thus by the number of limbs, i.e. by $k_2 \cdot \max r G + 1$.

(3) $\#\{i \mid 1 \leq i \leq p, w_i \in M_3\}$ is bounded by the number of true spikes of T , thus by $\# \Sigma \cdot (k_2 \cdot \max r G + 1)$ (see Lemma 4.2(1)).

(4) $\#\{i \mid 1 \leq i \leq p, w_i \in M_4\}$ is bounded by the number of true spikes of T , thus by $\# \Sigma \cdot (k_2 \cdot \max r G + 1)$ (see Lemma 4.2(1)).

Combining (1) through (4) we get that p is bounded by

$$(k_2 \cdot \max r G + 1)(\# \Sigma \cdot (\max r G + 2) + 1) = k.$$

Thus the lemma holds. ■

5. Contributions of Spikes.

In view of the results from the previous section, to prove that for each $\alpha \in \text{type 3II}$, $L(G_\alpha) \in CB$, it remains to prove that $LC(\beta) \in CB$ and $RC(\beta) \in CB$ for each $\beta \in \text{type 3II}$. This is done in the present section. Since the situation for left and right contributions is "symmetric", without loss of generality we consider right contributions only.

First we need the notion of a promised block. Intuitively a segment of a word is called a promised block if (as a whole) it will derive a block of a 's (a promised block of type A) or a block of b 's (a promised block of type B). Formally we have the following definition.

Definition 5.1. Let $x = a_1 \cdots a_n$, $a_i \in \text{us } G$ for $1 \leq i \leq n$, $(y, k, l) \in \text{SEG}(x)$.

Then y is called a *promised block of type A* if $x = a_1 \cdots a_{k-1} y a_{l+1} \cdots a_n$ and one of the following conditions holds.

- (1) $y \in (\text{type } 2b)(\text{type } 1a)^*(\text{type } 2a)$.
- (2) $y \in (\text{type } 2b)(\text{type } 1a)^*$ and $a_{l+1} \in (\text{type } 1b \cup \text{type } 2b)$.
- (3) $y \in (\text{type } 1a)^*(\text{type } 2a)$ and $a_{k-1} \in (\text{type } 1b \cup \text{type } 2a)$.
- (4) $y \in (\text{type } 1a)^+$, $a_{k-1} \in (\text{type } 1a \cup \text{type } 2a)$ and $a_{l+1} \in (\text{type } 1a \cup \text{type } 2b)$.

We call y a *promised block of type B* if $x = a_1 \cdots a_{k-1} y a_{l+1} \cdots a_n$ and one of the following conditions holds.

- (1) $y \in (\text{type } 2a)(\text{type } 1b)^*(\text{type } 2b)$.
- (2) $y \in (\text{type } 2a)(\text{type } 1b)^*$ and $a_{l+1} \in (\text{type } 1a \cup \text{type } 2a)$.
- (3) $y \in (\text{type } 1b)^*(\text{type } 2b)$ and $a_{k-1} \in (\text{type } 1a \cup \text{type } 2b)$.
- (4) $y \in (\text{type } 1b)^+$, $a_{k-1} \in (\text{type } 1a \cup \text{type } 2b)$ and $a_{l+1} \in (\text{type } 1a \cup \text{type } 2a)$.

If y is a promised block of type A or a promised block of type B then y is called a *promised block*. ■

Definition 5.1 is illustrated (in the case of a promised block of type A) by Figure 7.

We then have the following lemma.

Lemma 5.1. Let $a \in \text{type III}$ and let T be a successful derivation tree of G_a which contains a spike $\sigma = \langle\langle v_1, \dots, v_p \rangle\rangle$ and there exists a positive integer t such that either $lresult_i(T, \sigma)$ or $rresult_i(T, \sigma)$ contains as subword a promised block x . Then

(a) If x is of type A then there exists a $\beta \in \text{alph } x$ which is of type 1 or 2

such that for every $n > 0$, $\beta \xrightarrow[G]{n} w$, $w \in (us G)^+$ implies $\#_{\text{type } 1a} w \geq n$.

(b) If x is of type B then there exists a $\beta \in \text{alph } x$ which is of type 1 or 2

such that for every $n > 0$, $\beta \xrightarrow[G]{n} w$, $w \in (us G)^+$ implies $\#_{\text{type } 1b} w \geq n$.

Proof. We will prove (a); the proof of (b) is analogous. That (a) holds is proved by contradiction. Assume that (a) does not hold. Then from Section 3 we know that for every $\beta \in \text{alph } x$ which is of type 1 or 2 there exists a positive integer C_β such that for every $n > 0$ there exists a $w_n \in (us G)^+$ where $\beta \xrightarrow[G]{n} w_n$ and $\#_{\text{type } 1a} w_n \leq C_\beta$. Let

$$D = \max\{C_\beta \mid \beta \in \text{alph } x\} \cdot |x| \cdot \max G.$$

Let f be such that $K \in LB$ with parameter f . Without loss of generality we assume that x is a subword of $lresult_i(T, \sigma)$.

Let T' be the subtree with root v_1 where we delete

- (i) all descendants of v_{p+1} ,
- (ii) all nodes of set T , and
- (iii) all nodes of $lset_i(T, \sigma)$, $i > t$.

Let $T_1, \dots, T_{f(D)+1}$ be $f(D)+1$ disjoint copies of T . Let $v_1^1, \dots, v_1^{f(D)+1}$ be the nodes corresponding to v_1 and let $v_{p+1}^1, \dots, v_{p+1}^{f(D)+1}$ be the nodes corresponding to v_{p+1} . Let T'' be the tree which results from $T_1, \dots, T_{f(D)+1}$ by identifying v_1^{j+1} and v_{p+1}^j for $1 \leq j \leq f(D)+1$. Let T''' be the tree which results from T by removing all nodes of set T and by replacing the subtree rooted at v_1 by T'' .

Finally T''' is completed to a successful derivation tree of G_α as follows. Let $m = \text{height } T''' + 1$. In each leaf node v of T''' with $l(v) = \beta \in \text{alph } x$ which occurs in set T''' , $j < m$, append a tree representing a derivation

$$D_{\beta,j} : \beta \xrightarrow[g]{m-j} w_{\beta,j} \in (\text{us } G)^+ \text{ where } \#_{\text{type } 1a} w_{\beta,j} \leq C_\beta$$

In each leaf node v of T''' , with $l(v) = \beta \notin \text{alph } x$ which occurs in set T''' , $j < m$

append an arbitrary tree representing a derivation $D_{\beta,j} : \beta \xrightarrow[G]{m-j} w_{\beta,j} \in (\text{us } G)^+$.

Let \bar{T} denote the resulting tree. This situation is illustrated by Figure 8 (for the case of $f(D) = 2$).

Since G satisfies assumptions (1) through (6) of Section 3 the above construction implies the existence of a word $w \in K$ such that $\#BL^D(w) \geq f(D)+1$; a contradiction. Hence (a) holds. \blacksquare

We are now ready to prove that for a $\alpha \in \text{type } 3II$, $RC(\alpha)$ and $LC(\alpha)$ belong to CB .

Lemma 5.2. Let $\alpha \in \text{type } 3II$. Then $RC(\alpha) \in CB$ and $LC(\alpha) \in CB$.

Proof. Let $g : \mathbb{N}^+ \rightarrow \mathbb{N}^+$ be defined by $g(n) = (\text{maxr } G)^{n+4}$. We will show that $RC(\alpha) \in CB$ with parameters 4 and g . The proof for $LC(\alpha)$ is analogous.

Let $x \in RC(\alpha)$. Thus we have a situation as expressed by Definition 4.2. We will use the notation of Definition 4.2. We number of levels from T bottom-up (starting from 1) and we consider the covering of $BL^t(x)$ where t is a positive integer. The situation is depicted in Figure 9.

Let $x = x_1 \cdots x_q$, $q \geq 1$ where $x_i \in BL(x)$ for $1 \leq i \leq q$. Blocks x_1 and x_q are called *outside blocks* and blocks x_2, \dots, x_{q-1} are called *inside blocks*.

Let $1 \leq i \leq q$ be such a block. Let $p(x_i)$ be the number of the level on which a node v_j of σ lies such that v_j is an ancestor of *last* x_i but v_{j+1} is not an ancestor of *last* x_i . We consider inside blocks only.

Let x_i be such a block. Observe that $p(x_i) \geq p(x_{i-1})$. Now x_i is called

t-young if $p(x_i) \leq t+4$,

t-old if $p(x_{i-1}) \geq t+4$,

t-middle if $p(x_{i-1}) \leq t+4$ and $p(x_i) \geq t+4$.

Claim 5.1. If x_i is *t*-old, then $|x_i| \geq t+1$.

Proof. Let u_1 (u_2 respectively) be the word formed by all ancestors of nodes from x_i on the level $p(x_{i-1})-1$ ($p(x_{i-1})-2$ respectively). The situation is depicted in Figure 10.

All letters of u_1 are of one of the following types: $3l, 2a, 2b, 1a, 1b$, and consequently all letters of u_2 are of one of the following types: $2a, 2b, 1a, 1b$. Consequently u_2 is a promised block (see Definition 5.1) and thus by Lemma 5.1 u_2 contributes at least $p(x_{i-1})-3$ letters a (or letters b) to the block x_i . Since x_i is *t*-old, $p(x_{i-1}) \geq t+4$ and consequently $p(x_{i-1})-3 \geq t+1$; thus $|x_i| \geq t+1$. ■

Claim 5.2. The joint length of all *t*-young blocks is bounded by $(\max G)^{t+4}$; moreover, all *t*-young blocks are adjacent.

Proof. Both facts follow from the fact that to the left of a *t*-young block (different from the first inside block) there is always a *t*-young block and all of them are included in the contribution of the node in σ which is an ancestor of the last letter of the rightmost *t*-young block (and this node is on a level not higher than $t+4$). ■

Claim 5.3. Among x_2, \dots, x_{q-1} there is at most one t -middle block.

Proof. This follows immediately from Claim 5.2 and the observation that each block to the right of a t -middle block is t -old and each block to the left of a t -middle block is t -young. ■

We have now the following situation concerning blocks in x .

- all t -old blocks are outside $BL^t(w)$,

- there are at most three ("special") blocks to handle: x_1, x_q and the t -middle block, each of them has length at most t and can be covered by a segment of length $(\max G)^{t+4}$ if $\max G \geq 2$ (clearly, this can be assumed without loss of generality).

- all t -young blocks form a subword of length bounded by $(\max G)^{t+4}$.

Then clearly we can choose $X \subseteq \text{SEG}(w)$ such that

- (i) $\#X \leq 4$;
- (ii) for every $z \in X$, $|z| \leq (\max G)^{t+4} = g(t)$; and
- (iii) X covers $BL^t(w)$.

Thus indeed $RC(\alpha) \in CB$ with parameters 4 and g .

We are now ready to state the analogous of Theorem 3.1 for letter of type 3II.

Theorem 5.1. If $\alpha \in \text{type } 3II$, then $L(G_\alpha) \in CB$.

Proof. Follows immediately from Theorems 2.2, 2.3, 3.1 and Lemmas 4.4 and 5.2. ■

Finally we get the following theorem.

Theorem 5.2. $K \in CB$.

Proof. Clearly $K = L(G) \subseteq (\bigcup_{\alpha \in us G} L(G_\alpha) \cup M_0)^{\leq k}$, where k is a positive integer and M_0 is a finite language. Then Theorems 2.2, 2.3, 3.1 and 5.1 imply $K \in CB$. ■

We are now ready to prove the main result of the paper.

Theorem 5.3. Let Δ be an alphabet, let $\Pi = (\Delta_1, \Delta_2)$ be a binary partition of Δ and let $L \subseteq \Delta^*$ be an EOL language such that $L \in LB(\Pi)$. Then $L \in CB(\Pi)$.

Proof. Let L be as in the statement of the theorem. Let h be the homomorphism defined by $h(\alpha) = a$ if $\alpha \in \Delta_1$ and $h(\alpha) = b$ if $\alpha \in \Delta_2$ where a and b are two different symbols.

Then by Theorem 2.4, $h(L) \in LB(\{a\}, \{b\})$. Then Theorem 5.2 implies $h(L) \in CB(\{a\}, \{b\})$ and consequently, again applying Theorem 2.4 yields $L \in CB(\Pi)$. ■

6. Applications.

In this section we present examples to show the usefulness of Theorem 5.3 for proving that a language is not an EOL language.

Example 6.1.

$K = \{a^{i_1}b^{j_1}a^{i_2}b^{j_2}\dots a^{i_n}b^{j_n} \mid n \geq 1 \text{ and } i_l, j_l \geq 1 \text{ for } 1 \leq l \leq n\}$. It was proved in Section 2 that for $\Pi = (\{a\}, \{b\})$, $K \in LB(\Pi) - CB(\Pi)$. Consequently Theorem 5.3 implies $K \notin L(EOL)$. Observe that $K \in L(ETOL)$. \square

Example 6.2. Let G be a rewriting system with letters a and b , productions $a \rightarrow b^2$ and $b \rightarrow a^2$, axiom a and with the following rewriting rule.

Let $x = a_1 \dots a_n$, $n \geq 1$, $a_i \in \{a, b\}$ for $1 \leq i \leq n$, and let $y = y_1 \dots y_n$. Then $x \xrightarrow{G} y$ if either $a_i \xrightarrow{P} y_i$ for $1 \leq i \leq n$, or there exists a $1 \leq j \leq n$ such that $y_i = a_i$ and for $1 \leq i \leq n$, $i \neq j$, $a_i \xrightarrow{P} y_i$. Thus in a string either all letters are rewritten or all but one letter are rewritten.

The system G is a regular pattern grammar (see, e.g., [KR1]). We will now show that $K = L(G)$ (which consists of all strings derivable from the axiom) is not an EOL language. Observe that we are going to prove this without knowing any explicit expression for the language K .

Let $\Pi = (\{a\}, \{b\})$.

Claim 6.1. $K \in LB(\Pi)$ with parameter f where $f(1) = 1$ and for $m > 1$, $f(m) = f(\lfloor m/2 \rfloor) + 3$. ($\lfloor m/2 \rfloor$ denotes the integer k such that $2k \leq m$ and $2(k+1) > m$).

Proof. The proof goes by induction on the length of the blocks m .

(1) $m = 1$.

We must prove that for every $w \in K$, $\#BL_{\Pi}^{\dagger}(w) \leq 1$. This is proved by induction on the number of derivation steps needed to derive w .

If w is the axiom then $w = a$ and thus $\#BL_{\Pi}^1(a) = 1$. Assume that for any $w \in K$ which can be derived in less than or equal to n steps, $\#BL_{\Pi}^1(w) \leq 1$. Then let $a \xrightarrow{n+1} w_{n+1}$, i.e. $a \xrightarrow{n} w_n \xrightarrow{1} w_{n+1}$ and by induction $\#BL_{\Pi}^1(w_n) \leq 1$.

To obtain w_{n+1} either all occurrences of letters from w_n are rewritten or all but one occurrence of letters from w_n are rewritten. If all occurrences of letters from w_n are rewritten then the form of the productions implies that every block of w_{n+1} has length at least 2, hence $\#BL_{\Pi}^1(w_{n+1}) \leq 1$. If one occurrence of a letter is not rewritten, then this occurrence is the only possible candidate to belong to a block of length 1, hence $\#BL_{\Pi}^1(w_{n+1}) \leq 1$.

This concludes the proof for the case $m = 1$.

(2) Assume that for every $w \in K$, $\#BL_{\Pi}^k(w) \leq f(k)$ for $1 \leq k \leq m$. Then we prove that for every $w \in K$, $\#BL_{\Pi}^{m+1}(w) \leq f(m+1)$. This is proved by induction on the length of a derivation of w :

If w is the axiom, $\#BL_{\Pi}^{m+1}(w) \leq f(m+1)$ clearly holds. Assume that for any $w \in K$ such that $a \xrightarrow{\leq n} w$, $\#BL_{\Pi}^{m+1}(w) \leq f(m+1)$. Then let $a \xrightarrow{n} w_n \xrightarrow{1} w_{n+1}$. As in (1) there are two possible ways to derive w_{n+1} from w_n . If productions are applied to all occurrences of letters of w_n then every element of $BL_{\Pi}^{m+1}(w_{n+1})$ must come from an element of $BL_{\Pi}^{\lfloor (m+1)/2 \rfloor}(w_n)$. Thus

$$\#BL_{\Pi}^{m+1}(w_{n+1}) \leq \#BL_{\Pi}^{\lfloor (m+1)/2 \rfloor}(w_n) \leq f(\lfloor (m+1)/2 \rfloor) \leq f(m).$$

If one occurrence of a letter of w_n is not rewritten then

$$\#BL_{\Pi}^{m+1}(w_{n+1}) \leq \#BL_{\Pi}^{\lfloor (m+1)/2 \rfloor}(w_n) + 3 \leq f(\lfloor (m+1)/2 \rfloor) + 3 \leq f(m).$$

This concludes the proof of (2).

From (1) and (2) the claim follows. ■

Claim 6.2. $K \notin CB(\Pi)$.

Proof. The proof goes by contradiction. Assume that $K \in CB(\Pi)$ with parameters n_0 and g . We will derive a word w which violates the property.

Let $n > g(2^{n_0+1}) \cdot (n_0+1)$ and let $t = g(2^{n_0+1})$. To get w we first derive a^{2^n} and then proceed n_0+1 steps (using the second rewriting rule) according to the following scheme (for each step we have underlined the occurrence which is not rewritten in that step.)

$$\begin{aligned}
\underline{a}a^{2^n-1} &\Rightarrow ab^t \underline{bb}^{x_1} \\
&\Rightarrow b^2 a^{2t} b a^t \underline{aa}^{x_2} \\
&\Rightarrow a^4 b^{4t} a^2 b^{2t} ab^t \underline{bb}^{x_3} \\
&\Rightarrow b^8 a^{8t} b^4 a^{4t} b^2 a^{2t} b a^t \underline{aa}^{x_4} \\
&\Rightarrow \dots
\end{aligned}$$

If (n_0+1) is odd we get

$$w = a^{2^{n_0+1}} b^{(2^{n_0+1}) \cdot t} a^{2^{n_0}} b^{2^{n_0} \cdot t} \dots a^{2^2} b^{2^{2 \cdot t}} a^{2^1} b^{2^1 \cdot t} ab^t bb^{x_{n_0+1}}$$

and if (n_0+1) is even we get the same word with the roles of a and b interchanged.

Without loss of generality assume that n_0 is odd (n_0 is even is symmetric). Then all blocks of w consisting of b 's only are longer than t . Let $m_0 = 2^{n_0+1}$. All blocks of length at most m_0 consist only of a 's. However no two different elements of $BL_{\Pi}^{m_0}(w)$ can be covered by a segment of length at most $g(m_0) = t$. Thus if X covers $BL_{\Pi}^{m_0}(w)$ then $\#X > n_0$; a contradiction. This ends the proof of Claim 5.2. ■

Since $K \in LB(\Pi) - CB(\Pi)$, by Theorem 5.3, $K \notin L(EOL)$. ■

Acknowledgement.

The first and the second author gratefully acknowledge the financial support of NSF grant MCS 79-05245. The second and the third author are indebted to NFWO Belgium for supporting their research; all authors are indebted to H.C.M. Kleijn and the referee for useful comments on the first version of the paper.

References.

- [ER] A. Ehrenfeucht and G. Rozenberg, "The number of occurrences of letters versus their distribution in some EOL languages," *Inform. and Control* 26 (1975), 256-271.
- [EnR] J. Engelfriet and G. Rozenberg, "A translational theorem for the class of EOL languages," *Inform. and Control* 50 (1981), 175-183.
- [KR1] H.C.M. Kleijn and G. Rozenberg, "Context-free like restrictions on selective rewriting", University of Leiden, Technical report 80-19.
- [KR2] H.C.M. Kleijn and G. Rozenberg, "On the generative power of regular pattern grammars," *Acta Informatica* 20 (1983), 391-411.
- [RS] G. Rozenberg and A. Salomaa, *The mathematical theory of L systems*, 1980, Academic Press, New York.

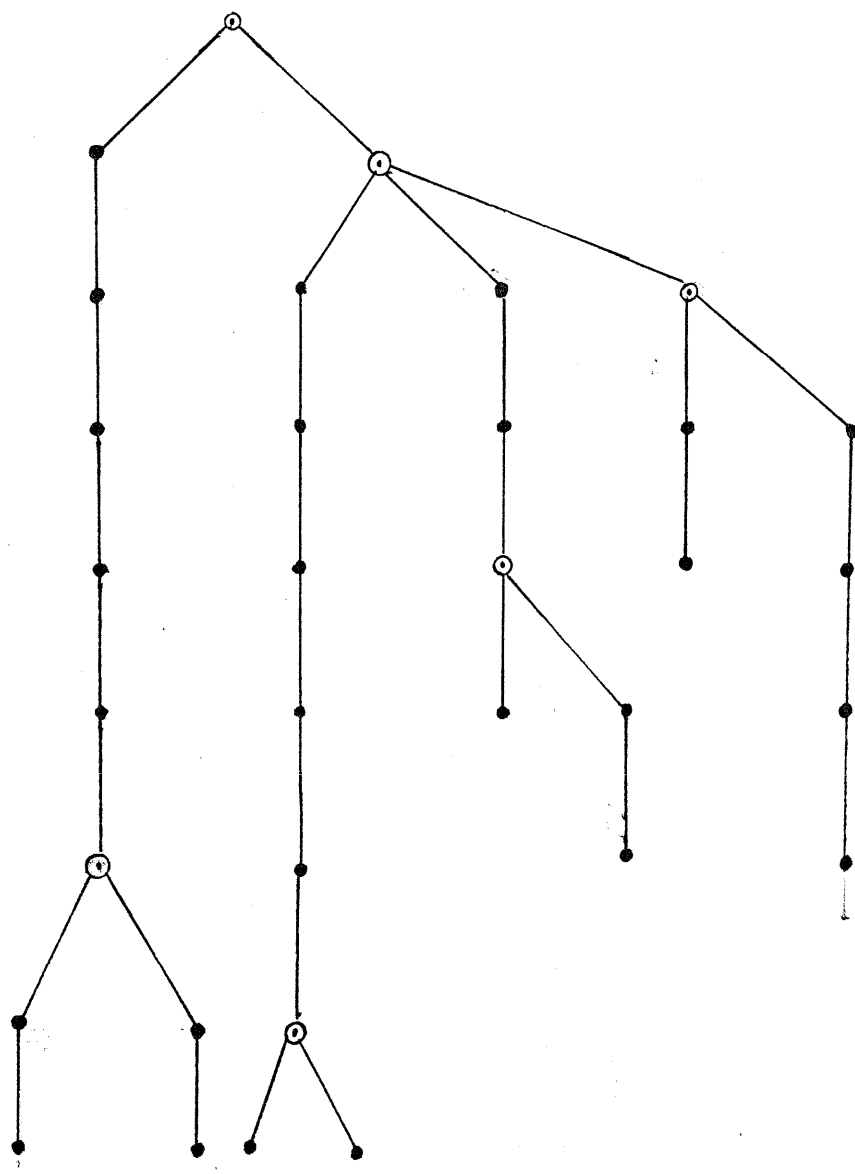


Figure 1

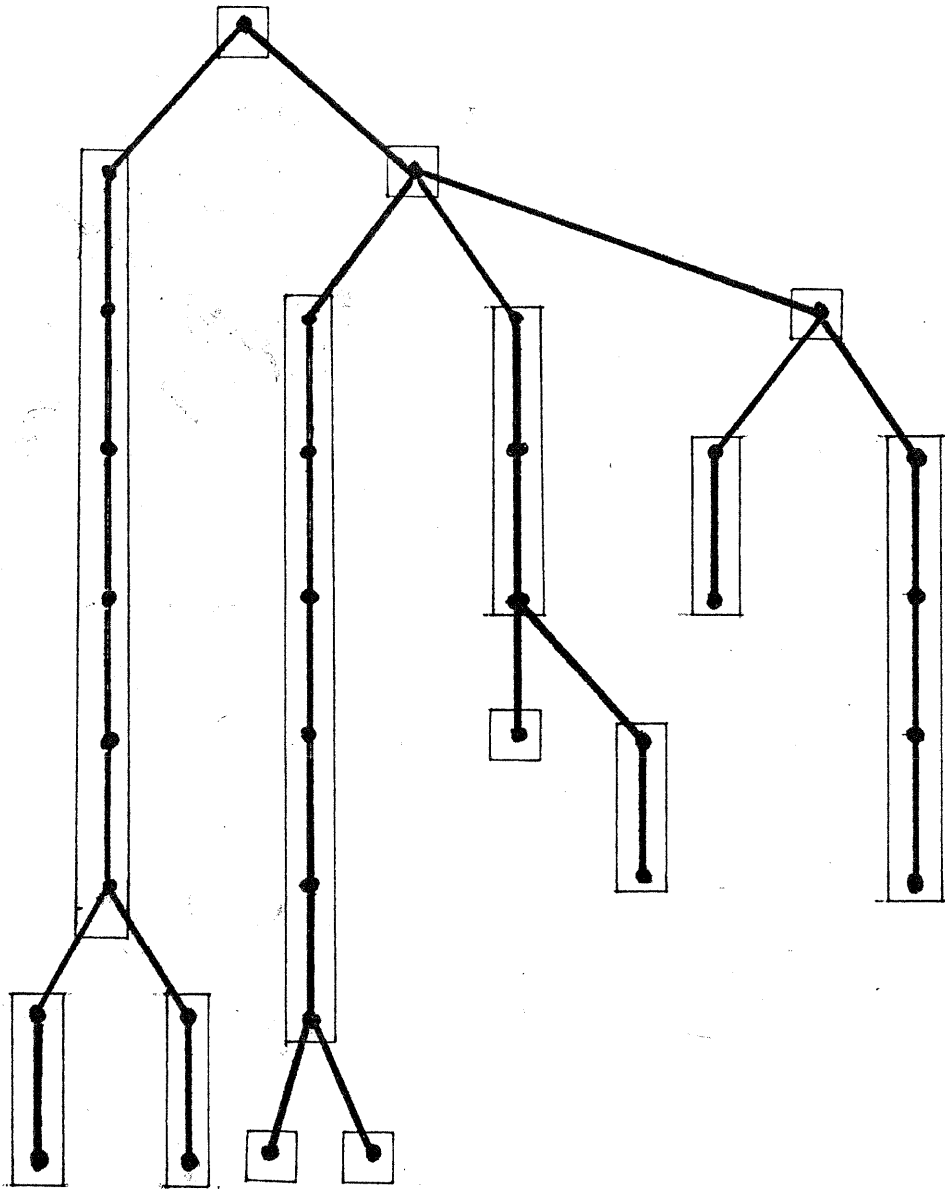


Figure 2

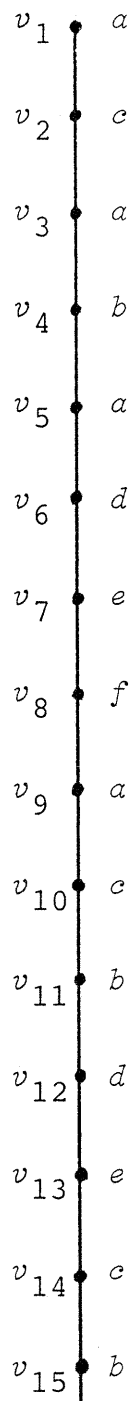


Figure 3

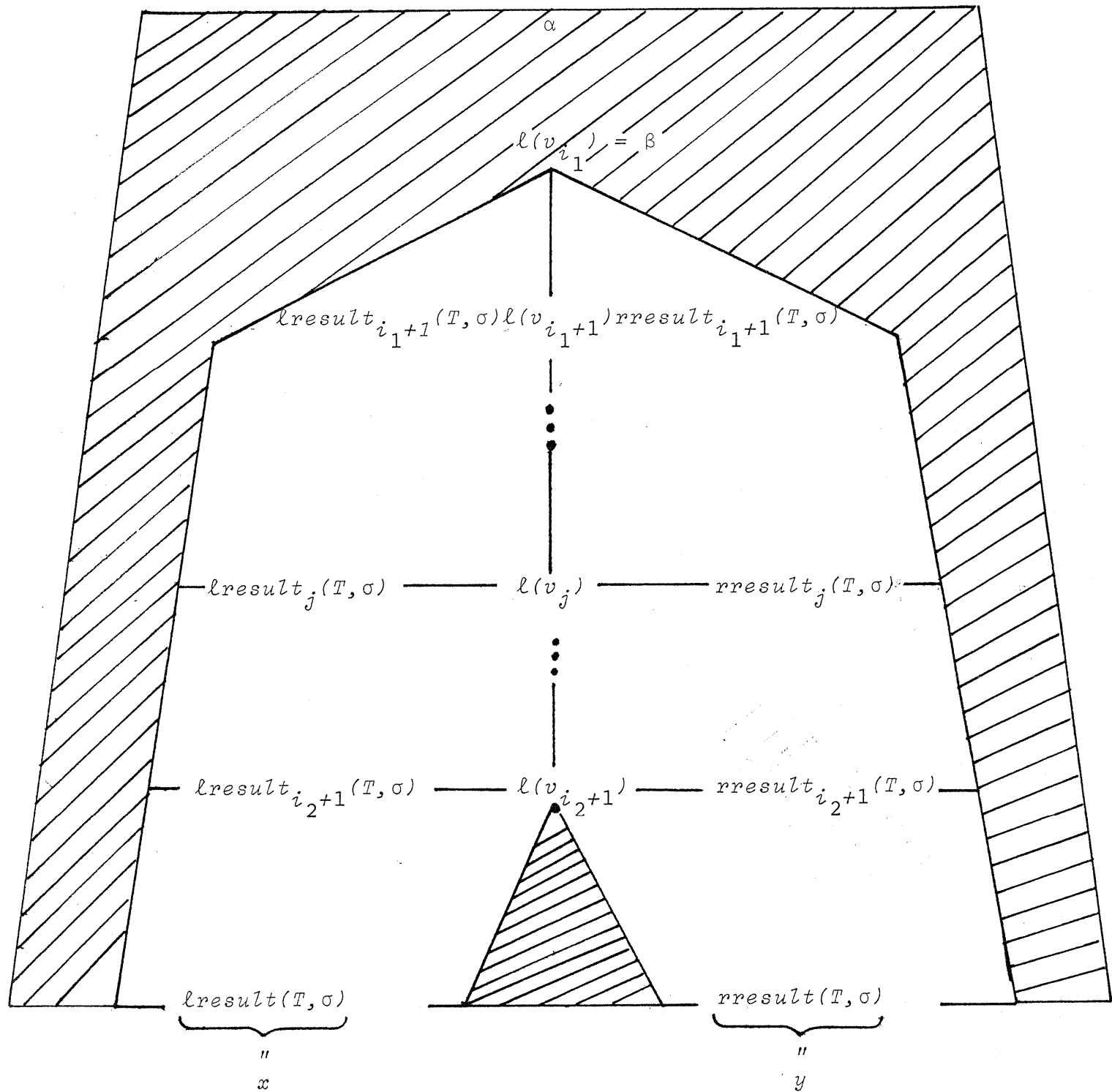


Figure 4

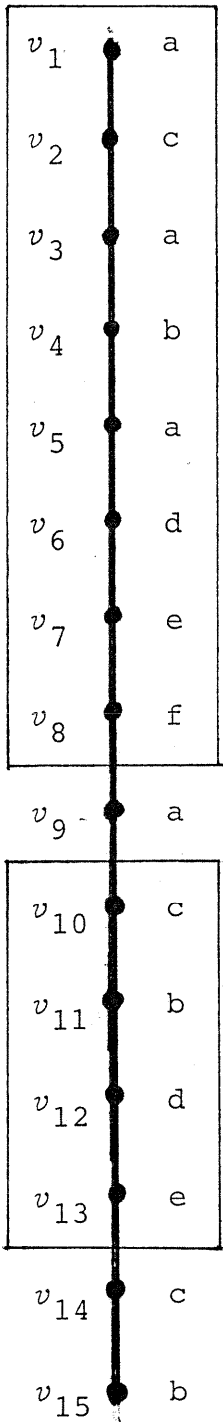


Figure 5

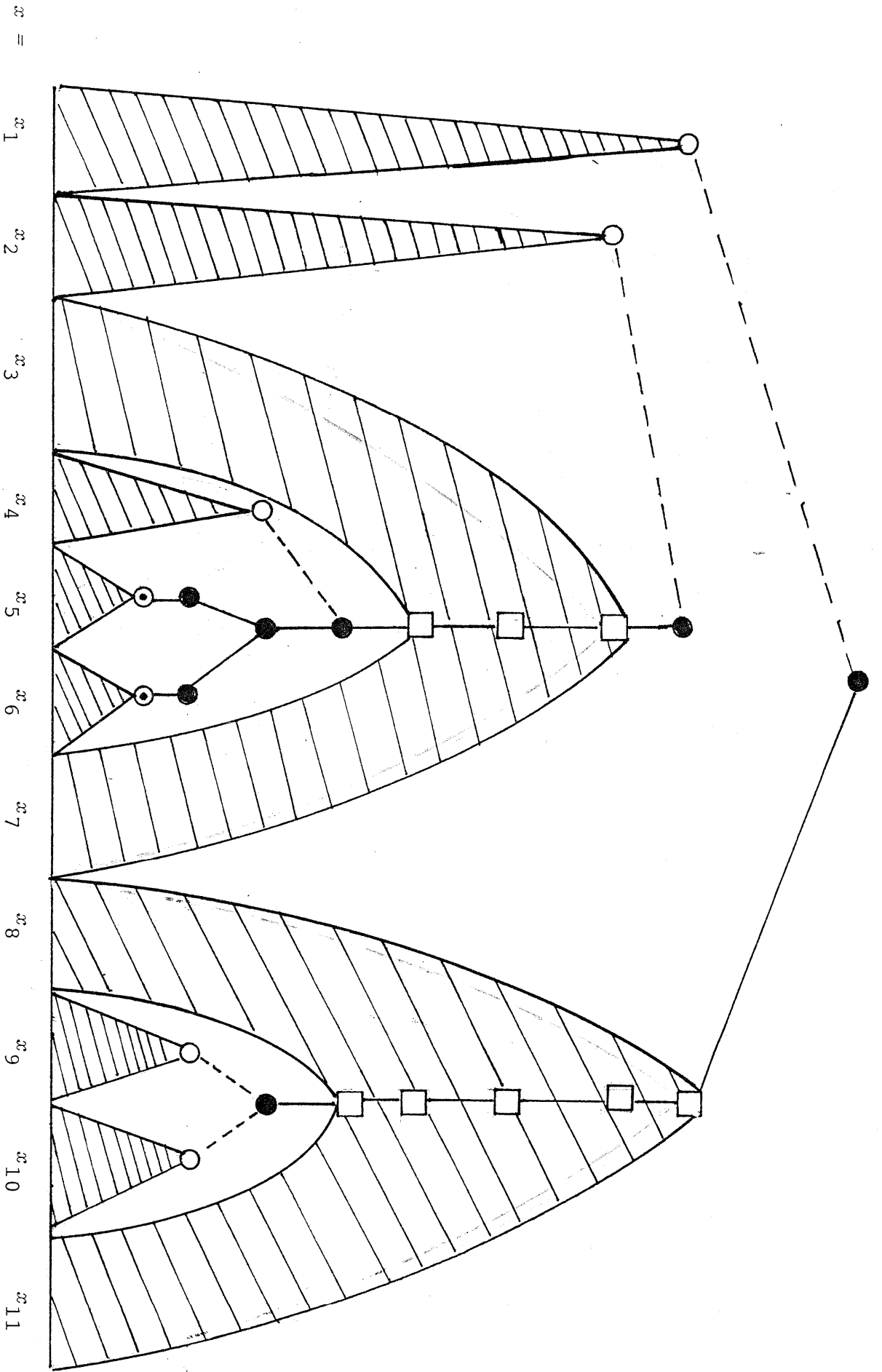


Figure 6

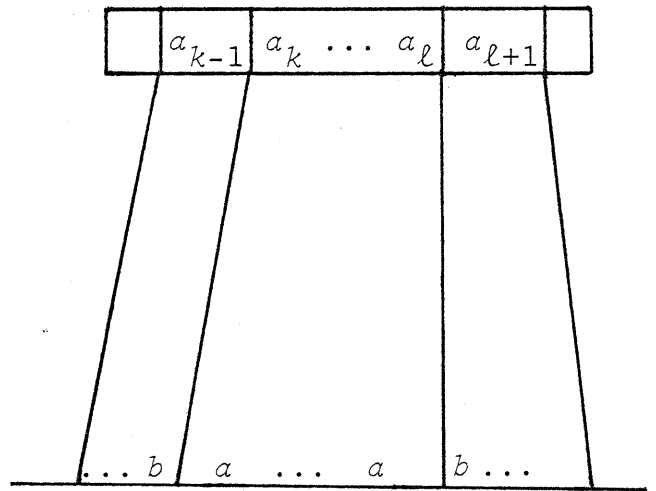
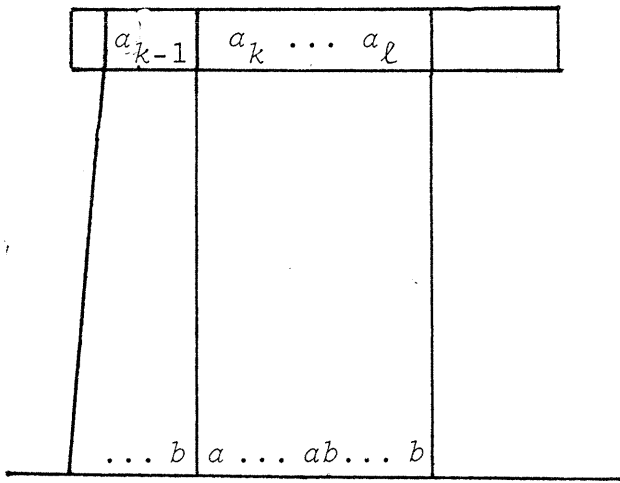
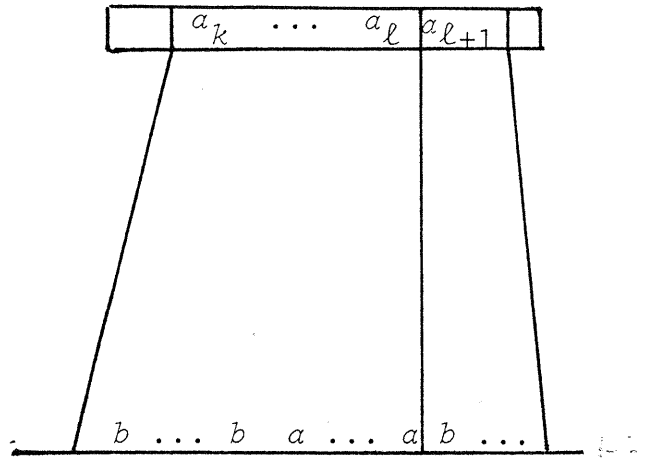
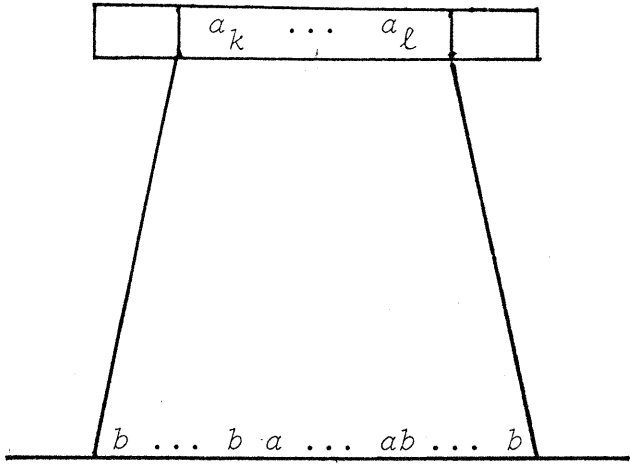
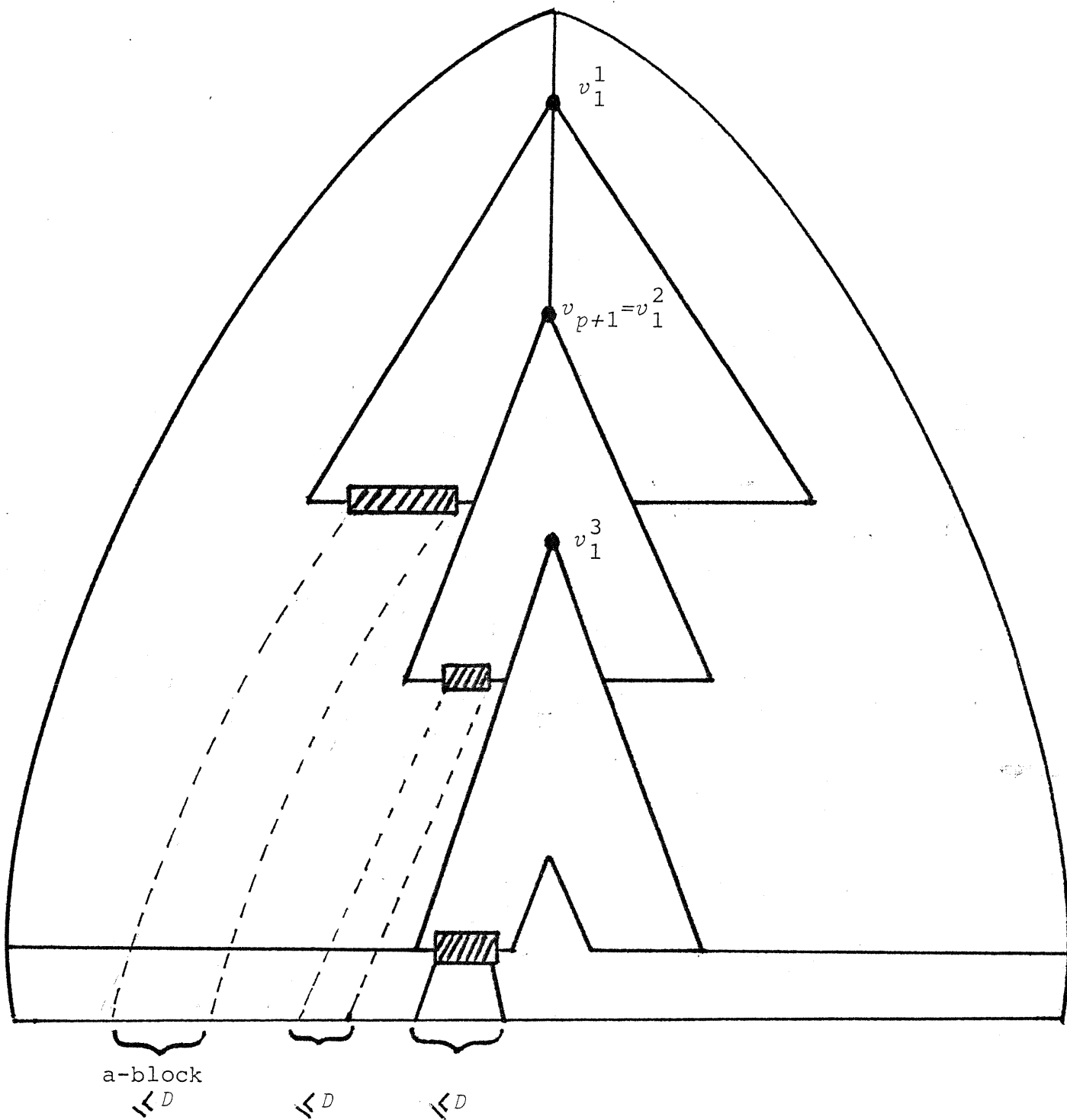


Figure 7




 denotes a promised block

Figure 8

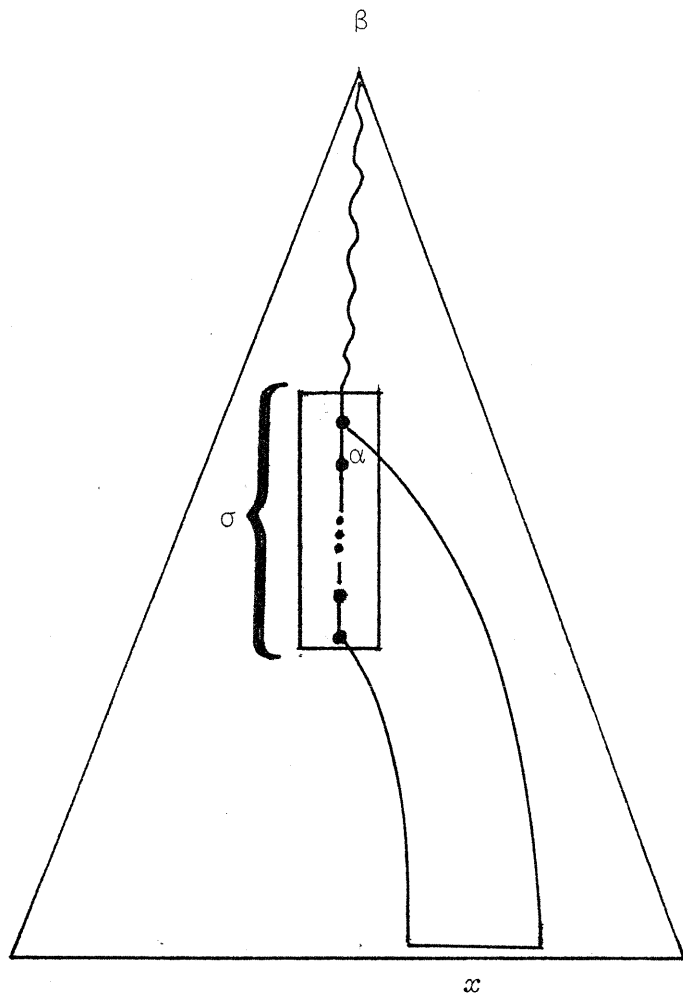


Figure 9

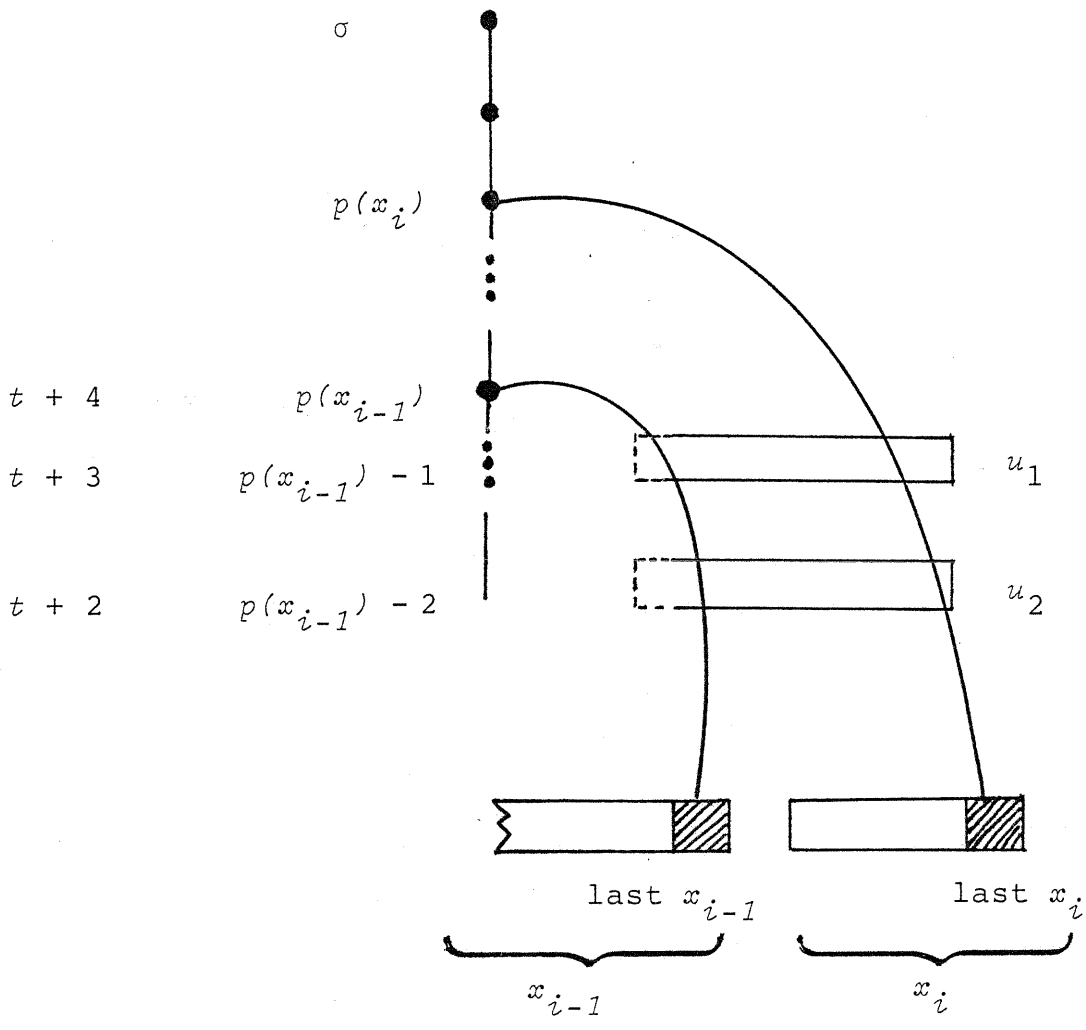


Figure 10