KYBERNETIKA — VOLUME *31* (1995), NUMBER 1, PAGES 83 – 97

ON THE GENERATIVE CAPACITY OF COLONIES

Gheorghe Păun[1]

We consider here colonies (grammar systems having as components regular grammars generating finite languages) with various derivation modes ($*, t, \leq k, = k, \geq k$, as usual in grammar systems area). Their generative capacity is investigated. Problems still open in the theory of general grammar systems (concerning, for instance, hierarchies on the number of components and on the parameter k mentioned above) are solved for this particular case. When hypothesis languages are added or the cooperation is aided by a transducer, the family of context-sensitive languages is characterized in most of these derivation modes.

1. INTRODUCTION

The notion of colony is introduced in [11] as an attempt to model in grammatical terms ideas in [1], [2], of considering intelligent systems build from as simple as possible elements. Informally speaking, a colony is a system of regular grammars, each generating a finite language, working on a common sentential form (in turn, each component replaces its axiom by a string), thus generating a language. Therefore we have a particular case of grammar systems, in the sense of [3], [4], a promising approach to distribution and cooperation appearing in various questions of artificial intelligence, cognitive psychology etc (see discussions in [10]). The colonies were investigated in [6], [12], [13] from various points of view, but a systematic study of them is still missing.

For instance, so far only the basic mode of derivation (one occurrence of a component axiom is replaced by a string) and the terminal mode (the maximal competence strategy: all occurrences of the axiom are replaced at a given step) have been considered (the last one in [13]). However, in grammar systems area more derivation modes are investigated: at least k, exactly k, at most k, any number of rules used when a component is enabled. In the case of colonies we do not count the used rules, but the number of axiom occurrences replaced by strings generated by the corresponding component. We start here the study of such variants, investigating their generative power. They look theoretically natural, but also "practically" motivated: think at intelligent systems whose components have to observe a precise protocol of cooperation, in terms of time restrictions about the work when enabled. More

[1] Research supported by the Alexander von Humboldt Foundation.

generally, the motivations for considering colonies with the basic mode of derivation [11] extend over these new derivation modes (consider, for example, a multi-agent system with time restrictions). However, our approach is basically mathematical, we do not look for specific "applications" of colonies with the working modes as described above.

Some results are as expected (hierarchies on the number of components), others are surprising (incomparability of families of languages defined by the parameter k, counting the number of replaced axioms). Both these problems are open for general grammar systems. It is also somewhat unexpected that colonies with further aid in work (hypothesis languages or sequential transducers) are as powerful as general grammar systems, characterizing again the family of context-sensitive languages.

2. COLONIES; DERIVATION MODES

For an alphabet V we denote by V^* the set of all strings of symbols in V, including the empty one, denoted by λ; as usual, $V^+ = V^* - \{\lambda\}$. The length of $x \in V^*$ is denoted by $|x|$ and $|x|_a$ denotes the number of occurrences of the symbol a in x.

We denote by FIN, REG, LIN, CF, CS the families of finite, regular, linear, context-free, context-sensitive languages, respectively (all languages are considered here modulo λ: two languages are identical if they differ at most in the empty string).

For all unexplained notions of formal language theory we refer to [16]; for regulated rewriting we refer to [7] and for grammar systems theory to [4].

A *colony*, in the sense of [11], is a construct

$$\sigma = (T, R_1, \ldots, R_n, w),$$

where T is an alphabet, $R_i = (N_i, T_i, P_i, S_i)$ are regular grammars with $L(R_i)$ finite, $1 \leq i \leq n$, and w is a string over $T \cup \{S_1, S_2, \ldots, S_n\}$ containing at least one symbol S_i, for some $i, 1 \leq i \leq n$; $T \subseteq \cup_{i=1}^{n} T_i$.

Because we are interested here only in the generative power of such machineries, we consider in the following an equivalent but more compact definition (also more suitable for introducing derivation modes of the types usual in grammar systems area).

Definition 1. A *colony* (of degree $n, n \geq 1$) is a construct

$$\sigma = (T, (S_1, F_1), \ldots, (S_n, F_n), w),$$

where T is an alphabet, $S_1, \ldots, S_n$ are symbols not in T (we denote $N = \{S_1, \ldots, S_n\}$), $F_i \subseteq (N \cup T - \{S_i\})^+$, $1 \leq i \leq n$, are finite languages, and $w \in (N \cup T)^* N (N \cup T)^*$.

Thus we are not interested in the way F_i is generated from S_i, but only in the strings it contains. Please note that in the style of [11], [12] we do not allow F_i to contain the empty string, and that the start string w contains at least one axiom symbol S_i.

Definition 2. For $x, y \in (N \cup T)^*,\ k \geq 1$, and a component (S_i, F_i) of a colony σ as above we define

$$\begin{aligned}
x \Longrightarrow_i^{=k} y \ \text{ iff } \ & x = x_1 S_i x_2 S_i \dots x_k S_i x_{k+1}, \\
& y = x_1 w_1 x_2 w_2 \dots x_k w_k x_{k+1}, \\
& \text{where } w_j \in F_i, 1 \leq j \leq k; \\
x \Longrightarrow_i^{\leq k} y \ \text{ iff } \ & x \Longrightarrow_i^{=k'} \text{ for some } k' \leq k; \\
x \Longrightarrow_i^{\geq k} y \ \text{ iff } \ & x \Longrightarrow_i^{=k'} \text{ for some } k' \geq k; \\
x \Longrightarrow_i^{*} y \ \text{ iff } \ & x \Longrightarrow_i^{=k} \text{ for some } k' \geq 0.
\end{aligned}$$

Moreover, we define

$$\begin{aligned}
x \Longrightarrow_i^{t} y \ \text{ iff } \ & x = x_1 S_i x_2 S_i \dots x_k S_i x_{k+1}, \\
& y = x_1 w_1 x_2 w_2 \dots x_k w_k x_{k+1}, \\
& \text{where } w_j \in F_i, 1 \leq j \leq k, \text{ and} \\
& |x_1 x_2 \dots x_{k+1}|_{S_i} = 0
\end{aligned}$$

(all occurrences of S_i are replaced by strings in F_i, not necessarily identical).

In [11], [12] only the derivation mode "$= 1$" is considered; the t-mode is investigated in [13].

Definition 3. The language generated by a colony σ in the derivation mode $f, f \in \{*, t\} \cup \{\leq k, = k, \geq k \mid k \geq 1\}$, is

$$\begin{aligned}
L_f(\sigma) = \{x \in T^* \mid \ & w \Longrightarrow_{i_1}^{f} w_1 \Longrightarrow_{i_2}^{f} \dots \Longrightarrow_{i_s}^{f} w_s = x, \\
& s \geq 1, 1 \leq i_j \leq n, 1 \leq j \leq s\}.
\end{aligned}$$

We denote by $\mathrm{COL}_n(f)$ the family of languages generated by colonies of degree at most $n, n \geq 1$, in the derivation mode f, f as above; we also put $\mathrm{COL}(f) = \bigcup_{n \geq 1} \mathrm{COL}_n(f)$.

3. THE GENERATIVE POWER

From definitions, we clearly have

$$\begin{aligned}
&\mathrm{COL}_1(f) \subseteq \mathrm{FIN}, \\
&\mathrm{COL}_n(f) \subseteq \mathrm{COL}_{n+1}(f),
\end{aligned}$$

for $n \geq 1,\ f$ as above, and the first relation is equality for $f \in \{*, t, = 1, \geq 1\} \cup \{\leq k \mid k \geq 1\}$. In [12] it is proved that

$$\mathrm{COL}(= 1) = \mathrm{CF} \quad \text{and} \quad \mathrm{COL}_n(= 1) \subset \mathrm{COL}_{n+1}(= 1),\ n \geq 1,$$

whereas in [13] it is proved that

$$\mathrm{CF} \subset \mathrm{COL}(t) = \mathrm{EPTOL}_{[1]},$$

where $\mathrm{EPTOL}_{[1]}$ is the family of languages generated by propagating ETOL systems having at most one rule $X \rightarrow x,\ x \neq X$, in each table (see [14]).

As clearly $L_{=1}(\sigma) = L_*(\sigma) = L_{\geq 1}(\sigma) = L_{\leq k}(\sigma),\ k \geq 1$, we have

Theorem 1. $\mathrm{COL}_n(=1) = \mathrm{COL}_n(*) = \mathrm{COL}_n(\leq k) = \mathrm{COL}_n(\geq 1) \subset \mathrm{CF}$, and $\mathrm{COL}(=1) = \mathrm{COL}(*) = \mathrm{COL}(\leq k) = \mathrm{COL}(\geq 1) = \mathrm{CF}$, for all $n \geq 1,\ k \geq 1$.

Consider now some *examples*: take the colony

$$\sigma_k = (\{a,b\}, (S_1, \{aS_2a, aba\}), (S_2, \{S_1\}), S_1^k),\ k \geq 1.$$

We have

$$L_{=k}(\sigma_k) = L_{\geq k}(\sigma_k) = \{(a^iba^i)^k \mid i \geq 1\},$$

a language which for $k \geq 2$ is not context-free.

Indeed, from a string containing less that k occurrences of the symbol S_1 or less than k occurrences of S_2 we cannot continue the derivation, hence the first component must either replace all occurrences of S_1 by aS_2a (and then the derivation continues) or all of them by aba (and the derivation is finished).

This is a finite index matrix language; consider also the colony (of degree 2)

$$\sigma = (\{a,b,c\}, (S_1, \{S_2S_2, S_2, aS_2b, ab\}), (S_2, \{S_1\}), (S_1c)^k).$$

It is easy to see that, for all $f \in \{*, t\} \cup \{\leq r \mid r \geq 1\} \cup \{= r, \geq r \mid 1 \leq r \leq k\},\ k \geq 1$,

$$L_f(\sigma) = \{x_1cx_2c\ldots x_kc \mid x_i \in D_{a,b} - \{\lambda\},\ 1 \leq i \leq k\},$$

where $D_{a,b}$ is the Dyck language over $\{a,b\}$ (the use of the replacement $S_1 \rightarrow S_2$ in the first component ensures the equality). Clearly, $L_f(\sigma)$ can be mapped into $D_{a,b}$ by a sequential transducer.

As $D_{a,b}$ is not a matrix language of finite index and the family of matrix languages of finite index (we denote it by $\mathrm{MAT}_{\mathrm{fin}}$) is closed under arbitrary sequential transducers [7], it follows that $L_f(\sigma)$ are not matrix languages of finite index.

The family $\mathrm{COL}_n(t),\ n \geq 3$, contains also one-letter non-regular languages: take

$$\sigma = (\{a\}, (S_1, \{S_2S_2\}), (S_2, \{S_1\}), (S_1, \{a\}), S_1).$$

Clearly,

$$L_t(\sigma) = \{a^{2^n} \mid n \geq 1\}.$$

On the other hand, as all the strings in colonies components are non-empty, a language $L_{=k}(\sigma)$ or $L_{\geq k}(\sigma)$ can contain only strings x with $|x| \geq k$. Therefore, for $k \geq 2$ and given V, languages $L \subseteq V^*$ containing strings $x \in V^*,\ |x| < k$, are not in $\mathrm{COL}(=k),\ \mathrm{COL}(\geq k)$, hence

Theorem 2. The families FIN, REG, LIN, CF, $\mathrm{MAT}_{\mathrm{fin}}$ are incomparable with each of $\mathrm{COL}_n(=k),\ \mathrm{COL}_n(\geq k),\ n \geq 2,\ \mathrm{COL}(=k),\ \mathrm{COL}(\geq k),\ k \geq 2$.

For the case $f = t,\ n = 2$ we have

Theorem 3. $\mathrm{COL}_2(t)$ is incomparable with $\mathrm{REG}, \mathrm{LIN}, \mathrm{MAT}_{\mathrm{fin}}$, but $\mathrm{COL}_2(t) \subset$ CF.

Proof. The inclusion $\mathrm{COL}_2(t) \subseteq \mathrm{CF}$ follows as in [3], [4] for context-free grammar systems, but, for technical reasons, we briefly prove it again: given $\sigma = (T, (S_1, F_1), (S_2, F_2), w)$ we construct the grammar

$$G = (\{S, S_1, S_2\}, T, \{S \to w\} \cup \{S_i \to x \mid x \in F_i,\ i = 1, 2\}, S).$$

In view of the fact that $F_i \subseteq (T \cup \{S_j\})^*,\ \{i, j\} = \{1, 2\}$, we have the equality $L_t(\sigma) = L(G)$.

This implies $\mathrm{Var}(L) \leq 3$ for every $L \in \mathrm{COL}_2(t)$ ($\mathrm{Var}(L)$ is the smallest number of nonterminals necessary in order to generate L by means of context-free grammars [8], [9]). As there are regular languages L with $Var(L) = n$ for every n, [8], we obtain $\mathrm{REG} - \mathrm{COL}_2(t) \neq \emptyset$.

On the other hand, as we have already pointed out, $\mathrm{COL}_2(t)$ contains languages not in $\mathrm{MAT}_{\mathrm{fin}}$. □

A natural and important problem is whether the parameters n and k, in $\mathrm{COL}_n(f)$, $f \in \{*, t\} \cup \{\leq k, = k, \geq k \mid k \geq 1\}$, lead to infinite hierarchies (are the colonies with $n+1$ components more powerful than those with n components ?). This is proved in [11] for $\mathrm{COL}_n(=1)$, hence also $\mathrm{COL}_n(*)$, $\mathrm{COL}_n(\geq 1)$, $\mathrm{COL}_n(\leq k)$, for given k, are infinite hierarchies, whereas $\mathrm{COL}_n(\leq k) = \mathrm{COL}_n(\leq 1)$, for all $k \geq 1$ and given n. The parameter n defines infinite hierarchies for the other modes of derivation too. (Note that this problem is still open for general grammar systems [4].)

Theorem 4. $\mathrm{COL}_n(t) \subset \mathrm{COL}_{n+1}(t),\ n \geq 1$.

Proof. For $n = 1, 2$ this is already proved ($\mathrm{COL}_1(t) = \mathrm{FIN} \subset \mathrm{COL}_2(t) \subset \mathrm{CF}$, but $\mathrm{COL}_3(t) - \mathrm{CF} \neq \emptyset$). Consider now the language

$$L_n = \bigcup_{i=1}^{n} (a^i b)^*,$$

for $n \geq 2$. It can be generated by the colony

$$\begin{aligned}\sigma_n = (\{a, b\}, (S_0, \{S_i \mid 1 \leq i \leq n\}),\quad & (S_1, \{abS_1', ab\}), (S_1', \{S_1\}),\\ & (S_2, \{a^2bS_2', a^2b\}), (S_2', \{S_2\}),\\ & \ldots\ldots\ldots\ldots\\ & (S_n, \{a^nbS_n', a^nb\}), (S_n', \{S_n\}), S_0).\end{aligned}$$

Therefore, $L_n \in \mathrm{COL}_{2n+1}(t)$.

Take now an arbitrary colony $\sigma = (\{a,b\},(S_1,F_1),\dots,(S_m,F_m),w)$ such that $L_t(\sigma) = L_n$. The derivations in σ can be viewed also as derivations in the context-free grammar $G = (\{S_0,S_1,S_2,\dots,S_m\},\{a,b\},\{S_0 \to w\} \cup \{S_i \to z \mid z \in F_i,\ 1 \leq i \leq m\},S_0)$ (but not conversely), hence we can speak about derivations in σ as derivations in G too. Because every subset $(a^ib)^*,\ 1 \leq i \leq n$, of L_n is infinite, we must have a recursive derivation $S_i \Longrightarrow^* uS_iv,\ uv \neq \lambda$; in order to obtain it, at least two components of σ are involved (S_i does not appear in strings of F_i). This derivation can be finished by some $S_i \Longrightarrow^* z$ and it can be iterated, $S_i \Longrightarrow^* u^rS_iv^r$, therefore u^rzv^r will eventually generate strings of arbitrarily large length (the colony is λ-free), hence containing substrings of the form ba^ib. Two such subderivations $S_i \Longrightarrow^* uS_iv$ must involve different nonterminal symbols (S_i, plus some $S_i', S_i'', \dots$ used in intermediate steps), otherwise strings containing two different substrings $ba^ib,\ ba^jb,\ i \neq j$, could be obtained. This implies that at least $2n$ different components are used in these n derivations $S_i \Longrightarrow^* uS_iv$. The symbols S_i must also be produced (possibly in strings in $(T \cup \{S_i\})^*, 1 \leq i \leq m$, not necessarily as strings S_i) by a further component, starting from the axiom of σ. In conclusion, σ must have at least $2n+1$ components, that is $L_n \notin \mathrm{COL}_{2n}(t)$. (The idea in the above sketched argument is used in many similar contexts in the descriptional complexity area – see, for instance, [8] – so we do not specify here all the technical details of the proof.)

We have obtained $\mathrm{COL}_{2n}(t) \subset \mathrm{COL}_{2n+1}(t),\ n \geq 2$.

For the language

$$L_n' = \bigcup_{i=2}^{n}(a^ib)^* \cup \{a^{2^i} \mid i \geq 1\},$$

$n \geq 2$, we have $L_n' = L_t(\sigma')$ for

$$\begin{aligned}\sigma' = (\{a,b\},(S_0,\{S_i \mid 1 \leq i \leq n\}),\quad & (S_1,\{S_1'S_1'\}),(S_1',\{S_1\}),(S_1,\{a\}),\\ & (S_2,\{a^2bS_2',a^2b\}),(S_2',\{S_2\}),\\ & \dots\dots\dots\dots\\ & (S_n,\{a^nbS_n',a^nb\}),(S_n',\{S_n\}),S_0),\end{aligned}$$

hence $L_n' \in \mathrm{COL}_{2n+2}(t)$. However, $L_n' \notin \mathrm{COL}_{2n+1}(t)$; the argument is the same as for the language L_n, with the remark that the sublanguage $\{a^{2^i} | i \geq 1\}$ of L_k' is not context-free, hence it requests at least three components to cooperate in producing it (see again Theorem 3).

It has remained the case $\mathrm{COL}_3(t) \subset \mathrm{COL}_4(t)$. For, consider the colony

$$\sigma = (\{a,b,c\},(S_0,\{bS_1,cS_1\}),(S_1,\{S_1'S_1'\}),(S_1',\{S_1\}),(S_1,\{a\}),S_0).$$

We have

$$L_t(\sigma) = \{ba^{2^n},ca^{2^n} \mid n \geq 1\}$$

and this language cannot be generated by a colony with only three components. (We need three components for producing $a^{2^n},\ n \geq 1$, but they cannot introduce also symbols b,c, because b,c appear only once each in the strings of $L_t(\sigma)$, hence they cannot appear in cycles and cannot be produced by replacing symbols which appear at least two times in the sentential forms). Thus $L_t(\sigma) \notin \mathrm{COL}_3(t)$, which completes the proof. □

A similar result can be obtained for the derivation modes $= k, \geq k$ for $k \geq 2$.

Theorem 5. $\mathrm{COL}_n(f) \subset \mathrm{COL}_{n+1}(f),\ n \geq 1,\ f \in \{= k, \geq k \mid k \geq 2\}$.

Proof. Consider the colony

$$\begin{aligned}\sigma_n = (\{a,b,c\},\ & (S_1, \{abS_1'ab, abcab\}), (S_1', \{S_1\}),\\ & (S_2, \{a^2bS_2'a^2b, a^2bca^2b\}), (S_2', \{S_2\}),\\ & \ldots\ldots\ldots\ldots\ldots\\ & (S_n, \{a^nbS_n'a^nb, a^nbca^nb\}), (S_n', \{S_n\}), (S_1S_2\ldots S_n)^k),\end{aligned}$$

for $n \geq 1$. We obtain

$$\begin{aligned}L_{=k}(\sigma_n) &= L_{\geq k}(\sigma_n) =\\ &= \{((ab)^{i_1}c(ab)^{i_1}(a^2b)^{i_2}c(a^2b)^{i_2}\ldots(a^nb)^{i_n}c(a^nb)^{i_n})^k \mid\\ &\quad i_j \geq 1,\ 1 \leq j \leq n\}.\end{aligned}$$

Take now a colony $\sigma = (\{a,b,c\}, (S_1,F_1), \ldots, (S_m,F_m), w)$ such that $L_{=k}(\sigma) = L_{=k}(\sigma_n)$. Consider a substring $(a^rb)^{i_r}c(a^rb)^{i_r}$ of a string in $L_{=k}(\sigma_n)$, $1 \leq r \leq n$. As i_r can be arbitrarily large, in generating such a substring at least one cycle in σ is involved, $S_r \Longrightarrow^* uS_rv$, for some strings u, v eventually generating strings containing substrings $(a^rb)^i,\ i \geq 1$. The symbol c cannot appear in such u, v or in strings they generate (otherwise an unbounded number of c occurrences will be produced). Therefore c is introduced only in terminal non-recurrent derivations, $S_r \Longrightarrow^* xcy$. If either $u = \lambda$ or $v = \lambda$, then strings containing $(a^rb)^ic(a^rb)^j,\ i \neq j$, are obtained, a contradiction. Similarly when either u or v eventually generates strings containing substrings $(a^sb)^i$ with $s \neq r$. (Having k occurrences of S_r in some sentential form, using the subderivation $S_r \Longrightarrow^* uS_rv$ for all positions, the number of nonterminal occurrences remain multiple of k for every $S_i,\ 1 \leq i \leq m$, hence this derivation can be correctly ended.) Finally, if the strings u, v above will eventually produce terminal strings containing both $(a^rb)^i, i \geq 1$, and $(a^sb)^j,\ j \geq 1$, for some $r \neq s$, then by iterating this derivation we can get strings containing arbitrarily many substrings $(a^rb)^i$ in the right of which substrings $(a^sb)^j$ appear and conversely; such strings are not in $L_{=k}(\sigma_n)$ (in the strings of this language we have exactly n substrings $(a^tb)^{i_t}c(a^tb)^{i_t}$ for each $t,\ 1 \leq t \leq n$). Consequently, for each derivation $S_r \Longrightarrow^* uS_rv$ we get a terminal string $u'xcyv'$ with $u \Longrightarrow^* u',\ v \Longrightarrow^* v', S_r \Longrightarrow^* xcy$ such that $u'xcyv' = z_1(a^rb)^ic(a^rb)^jz_2,\ i, j \geq 1$, possibly with z_1 a non-empty suffix of a^rb and z_2 a non-empty prefix of a^rb and given difference $i - j$, possibly not zero. However, by iterating the subderivation $S_r \Longrightarrow^* uS_rv$ we can assume that i and j are arbitrarily large.

It is now clear that for two such derivations $S_r \Longrightarrow^* uS_rv, S_s \Longrightarrow^* u'S_sv'$ we cannot have $S_r = S_s$ (the two derivations can then be mixed and one generates strings with arbitrarily many substrings $(a^rb)^i$ followed by $(a^sb)^j$ and conversely). Every such cycle must involve two nonterminals, hence two components of σ. In conclusion, σ must contain at least $2n$ components, which implies $L_{=k}(\sigma_n) \notin \mathrm{COL}_{2n-1}(=k)$.

The same argument shows that $L_{\geq k}(\sigma_n) \notin \mathrm{COL}_{2n-1}(\geq k)$.

We have thus obtained the proper inclusions $\mathrm{COL}_{2n-1}(f) \subset \mathrm{COL}_{2n}(f)$, $n \geq 1$, $f \in \{=k, \geq k \mid k \geq 2\}$.

Modify now the above colony σ_n as follows:

$$\begin{aligned}\sigma'_n = (\{a,b,c\}, \quad & (S_0, \{c^2 S_1 S_2 \ldots S_n, c^3 S_1 S_2 \ldots S_n\}),\\ & (S_1, \{abS'_1ab, abcab\}), (S'_1, \{S_1\}),\\ & (S_2, \{a^2bS'_2a^2b, a^2bca^2b\}), (S'_2, \{S_2\}),\\ & \ldots\ldots\ldots\ldots\ldots\\ & (S_n, \{a^nbS'_na^nb, a^nbca^nb\}), (S'_n, \{S_n\}), S_0^k).\end{aligned}$$

It is easy to see that we need now $2n$ components in order to obtain recurrent derivations associated to the n substrings of the form $(a^rb)^ic(a^rb)^i$ and one more component introducing k times substrings c^2 and c^3 (such substrings cannot be produced by symbols involved in cycles, as these symbols will either generate arbitrarily many substrings c^2, c^3, or they will introduce such substrings only in the middle of the generated strings). In conclusion, $L_{=k}(\sigma'_n) = L_{\geq k}(\sigma'_n) \in \mathrm{COL}_{2n+1}(f) - \mathrm{COL}_{2n}(f)$, $n \geq 1$, $f \in \{=k, \geq k \mid k \geq 2\}$, which completes the proof. □

Somewhat surprising, the parameter k in derivation modes $=k, \geq k$ does not give infinite hierarchies, but infinitely many incomparable families. In order to prove this, we shall use the next pumping property.

Theorem 6. If $L \subset V^*, L \in \mathrm{COL}(=k)$ (or $L \in \mathrm{COL}(\geq k)$), $k \geq 1$, is an infinite language, then there is a string $y \in L$, which can be written in the form

$$y = u_1vw_1xu_2vw_2x \ldots u_kvw_kxu_{k+1},$$

with $u_i, w_i \in V^*$ for all i, $vx \neq \lambda$, and

$$u_1v^jw_1x^ju_2v^jw_2x^j \ldots u_kv^jw_kx^ju_{k+1}$$

is in L for all $j \geq 1$.

Proof. Take a colony $\sigma = (T, (S_1, F_1), \ldots, (S_n, F_n), z_0)$ with infinite $L_f(\sigma)$, $f \in \{=k, \geq k \mid k \geq 1\}$. Consider an arbitrary derivation in σ,

$$D : z_0 \Longrightarrow_{i_1}^{f} z_1 \Longrightarrow_{i_2}^{f} \ldots \Longrightarrow_{i_r}^{f} z_r = z \in L_f(\sigma).$$

This derivation can be completed with an initial step $S \Longrightarrow z_0$ and thus we have a context-free derivation in the grammar

$$G = (\{S, S_1, \ldots, S_n\}, T, \{S \to z_0\} \cup \{S_i \to y \mid y \in F_i,\ 1 \leq i \leq n\}, S).$$

Therefore we can consider the derivation tree associated to it in the usual way.

As $L_f(\sigma)$ is infinite, we can take z arbitrarily long, hence we may assume that the above considered derivation tree contains a path from S to the leafs with two nodes marked by the same symbol S_i. This implies that a subderivation $S_i \Longrightarrow^* v_0 S_i x_0$ there is in D using the components $i, i'_1, \dots,\ i'_m, m \geq 1$. If all such subderivations have $vx = \lambda$, for $v_0 \Longrightarrow^* v, x_0 \Longrightarrow^* x, v, x \in T^*$, subderivations in D, then we cannot obtain z of arbitrarily large length (the set of non-recurrent derivations is finite).

Look now at the strings $z_t, z_s,\ t < s$, where the two occurrences of S_i in $S_i \Longrightarrow^* v_0 S_i x_0$ appear, namely $z_t = z'_t S_i z''_t, z_s = z'_s v_0 S_i x_0 z''_s$. In the step $z_t \Longrightarrow_i^f z_{t+1}$ exactly k (at least k in the $\geq k$ mode of derivation) occurrences of S_i are rewritten. Choose k of them (hence in both modes of derivation we proceed in the same way), $z_t = \alpha_1 S_i \alpha_2 S_i \dots \alpha_k S_i \alpha_{k+1}$, and continue the derivation by using the rules in $S_i \Longrightarrow^* v_0 S_i x_0$ for all these k occurrences of S_i. (This is clearly possible.) In this way we obtain a derivation

$$\begin{aligned} z_0 \Longrightarrow_{i_1}^f z_1 \Longrightarrow_{i_2}^f \dots \Longrightarrow_{i_t}^f z_t &= \alpha_1 S_i \alpha_2 \dots \alpha_k S_i \alpha_{k+1} \\ &\Longrightarrow_i^f z'_{t+1} \Longrightarrow_{i'_1}^f z'_1 \Longrightarrow_{i'_2}^f \dots \Longrightarrow_{i'_m}^f z'_m = \\ &= \alpha_1 v_0 S_i x_0 \alpha_2 \dots \alpha_k v_0 S_i x_0 \alpha_{k+1}. \end{aligned}$$

These steps involving $S_i \Longrightarrow^* v_0 S_i x_0$ can be iterated $j \geq 1$ times:

$$\dots \Longrightarrow^* \alpha_1 v_0^j S_i x_0^j \alpha_2 \dots \alpha_k v_0^j S_i x_0^j \alpha_{k+1} = z'.$$

All the strings $\alpha_1, \dots, \alpha_{k+1}$ appear in z_t, together with k occurrences of S_i, therefore we can derive them exactly as in D, until obtaining terminal strings, $\alpha_q \Longrightarrow^* u_q,\ 1 \leq q \leq k+1, S_q \Longrightarrow^* w_q,\ 1 \leq q \leq k$.

In the string z' we have exactly k occurrences of v_0^j and also k occurrences of x_0^j. Irrespective how many nonterminals appear in v_0, x_0, their number is a multiple of k. Consequently, a terminal correct derivation in σ can be obtained (we choose k occurrences of some nonterminal S_h and replace each occurrence by the same string $y \in F_h$; continuing in this way, the number of nonterminals remains multiple of k, hence the derivation can be finished). Write $v_0 \Longrightarrow^* v, x_0 \Longrightarrow^* x$. In conclusion, we eventually obtain a string of the form

$$u_1 v^j w_1 x^j u_2 v^j w_2 x^j \dots u_k v^j w_k x^j u_{k+1},$$

which completes the proof (take as y the string as above with $j = 1$). □

This theorem has a series of important consequences.

Corollary 1. All families COL($= k_1$), COL($= k_2$), COL($\geq k_3$), COL($\geq k_4$) with different k_1, k_2, k_3, k_4 are pairwise incomparable.

Proof. The language

$$L = \{(a^n b^n)^k \mid n \geq 1\}$$

is in COL($= k$) $\cap$ COL($\geq k$), but not in COL($= j$) $\cup$ COL($\geq j'$) for $j \neq k \neq j'$ (use the previous necessary condition). □

Corollary 2. The length set of languages in COL($= k$), COL($\geq k$), $k \geq 1$, contains infinite arithmetical progressions.

Proof. Directly from Theorem 6. □

Theorem 7. The family COL(t) is incomparable with each of COL(=k), COL($\geq k$), $k \geq 2$.

Proof. The language $\{a^{2^n} | n \geq 1\}$ is in COL(t), but it does not have the property in Corollary 2. On the other hand, consider the colony

$$\sigma = (\{a,b\}, (S_1, \{aS_2b, ab\}), (S_2, \{S_1\}), S_1^{2k}).$$

Both $L_{=k}(\sigma)$ and $L_{\geq k}(\sigma)$ (in fact, $L_{=k}(\sigma) \subseteq L_{\geq k}(\sigma)$) contain strings of the form

$$a^{n_1}b^{n_1}a^{n_2}b^{n_2} \ldots a^{n_{2k}}b^{n_{2k}},$$

with the following two *properties*:

1. for every $i \neq j$, the difference $n_i - n_j$ can be arbitrarily large (start from S_1^{2k} and construct a derivation consisting of subderivations which rewrite different k occurrences of S_1 into aS_2b and back to S_1, iterated, in such a way that the i-th substring a^nb^n is arbitrarily increased but the j-th substring is bounded; we can do this because we have $2k$ nonterminal occurrences);
2. for no $j, 1 \leq j \leq 2k$, we can have n_j arbitrarily large and all $n_s, s \neq j$, bounded by a given constant (we have to increase at the same time the length of at least k subwords a^nb^n).

Suppose now that $L_f(\sigma) \in$ COL(t), $L_f(\sigma) = L_t(\sigma')$, for some colony $\sigma' = (\{a,b\}, (S_1, F_1), \ldots, (S_m, F_m), w)$, $f \in \{= k, \geq k\}$.

In view of the above properties of strings of the considered forms, we must have in σ' derivations of the form $S_i \Longrightarrow^* a^rS_ib^r$, $r \geq 1$ (we cannot have separate derivations $S_i \Longrightarrow^* a^rS_i, S_i \Longrightarrow^* S_ib^r$, because then the equality of the number of occurrences of a and b cannot be observed; derivations of different forms will mix the symbols a and b – consider, for instance, $S_i \Longrightarrow^* a^rS_i, S_i \Longrightarrow^* b^rS_i$; the same for derivations of other forms).

We consider from now on only the components (S_i, F_i) corresponding to symbols S_i involved in such recurrent derivations. Denote their set by M.

In each sentential form of σ' there are at most $2k$ occurrences of nonterminals S_i, but every nonterminal S_i, if appearing in a sentential form, then it appears at least twice (otherwise the previous property 2 is violated).

The strings in these sets F_i must be of the forms $a^sS_jb^s$, $s \geq 0$, $j \neq i$, or $a^{s'}b^{s''}$, $s', s'' \geq 0$. If $s = 0$, then the symbols a, b in $S_i \Longrightarrow^* a^rS_ib^r$ are introduced in another component, F_ℓ, which is also in M. If F_i contains both a terminal and a nonterminal string, then we can replace all occurrences of S_i but one by a terminal string and the remained occurrences by a nonterminal string. In this way a derivation increasing only one substring a^nb^n is obtained, violating again property 2. Similarly if F_i contains two strings $a^sS_jb^s$, $a^tS_lb^t$, $j \neq \ell$ (we fix all occurrences of S_i but one by replacing them by one string and work arbitrarily many steps on the non-fixed component using the other string).

Consequently, all the sets F_i corresponding to components in M are singletons. This implies that each occurrence of S_i will lead to the same string $a^m S_i b^m$, m arbitrarily large. Finally, S_i is replaced by a terminal string, which possibly differ from an occurrence of S_i to another, but only in the range of finitely many possibilities of derivations without cycles. As we have pointed out, S_i associated to components in M appears at least twice in each sentential form. In this way, property 1 is violated (we need arbitrarily different powers for all substrings $a^{n_i} b^{n_i}$). Contradiction, hence we cannot have $L_f(\sigma) = L_t(\sigma')$. □

4. COLONIES WITH HYPOTHESIS LANGUAGES

A colony is a model of an intelligent system designed for solving certain problem, hence it is supposed that the system has a target, its actions tend to some expected results. This can be captured in colony terms by considering a target language, for selecting the sentential forms generated by the colony. For grammar systems this has been done in [5], by considering a regular language as hypothesis language of the system behavior. The result is that such systems (with context-free components) characterize the family of context-sensitive languages (this fits with results in regulated rewriting area [7], where similar characterizations of CS are obtained). We shall see here that this is true also for colonies, thus strenghtening the results in [5].

Definition 4. A *colony with (regular) hypothesis language* (a h-colony, for short) is a construct

$$\sigma = (T, (S_1, F_1), \ldots, (S_n, F_n), R, w),$$

where $(T, (S_1, F_1), \ldots, (S_n, F_n), w)$ is a usual colony and R is a regular language in $(N \cup T)^* - T^*$.

Now, for $f \in \{*, t\} \cup \{\leq k, = k, \geq k \mid k \geq 1\}$ and $1 \leq i \leq n$, we accept a derivation $x \Longrightarrow_i^f y$ only when $y \in R$ or $y \in T^*$. The generated language is

$$\begin{aligned} L_f(\sigma) = \{x \in T^* \quad | \; & w \Longrightarrow_{i_1}^f w_1 \Longrightarrow_{i_2}^f \ldots \Longrightarrow_{i_s}^f w_s = x, \\ & s \geq 1,\ 1 \leq i_j \leq n,\ 1 \leq j \leq s, \text{ and} \\ & w_j \in R,\ 1 \leq j \leq s-1\}. \end{aligned}$$

(No hypothesis is made about the last string, the terminal one.)

We denote by $\mathrm{HCOL}_n(f)$, $\mathrm{HCOL}(f)$, $n \geq 1, f$ as above, the families of languages obtained in this way, corresponding to $\mathrm{COL}_n(f)$, $\mathrm{COL}(f)$, respectively.

Every colony is a h-colony: take $R = (N \cup T)^* - T^*$, hence imposing no restriction. Therefore,

$$\mathrm{COL}_n(f) \subseteq \mathrm{HCOL}_n(f),\ n \geq 1,$$
$$\mathrm{COL}(f) \subseteq \mathrm{HCOL}(f),\ f \text{ as above}.$$

For $f \in \{= k, \geq k \mid k \geq 2\}$ we again have finite languages not in $\mathrm{HCOL}(f)$, but the hypothesis languages increase considerably the power of colonies of all types.

Theorem 8. $\mathrm{HCOL}(f) = \mathrm{CS},\ f \in \{*, t, =1, \geq 1\} \cup \{\leq k \mid k \geq 1\}$.

Proof. The inclusions $\subseteq$ are obvious (they can be obtained by straightforward constructions).

Conversely, take a grammar $G = (N, T, P, S)$ in Kuroda normal form, that is with rules in P of the forms

$$A \to a,\ A \to BC,\ AB \to CD,\ \text{for } A, B, C, D \in N,\ a \in T.$$

We allow also rules $A \to B,\ A,\ B \in N$. Without loss of generality we may assume that no lefthand nonterminal of a rule appears also in the righthand side of the same rule.

Assume $P = P_1 \cup P_2$ with

$$\begin{aligned} P_1 &= \{p_i : A \to x \mid 1 \leq i \leq n\}, \\ P_2 &= \{q_i : AB \to CD \mid 1 \leq i \leq m\}. \end{aligned}$$

We construct now a h-colony for the language $L(G)$. All the cases $f \in \{*, =1, \geq 1\}$ $\cup \{\leq k \mid k \geq 1\}$ can be treated together. Namely, take the colony σ with the terminal alphabet T, the axiom string S and the next components:

1. $(A, \{x\})$, for $p_i : A \to x \in P_1,\ 1 \leq i \leq n$,
2. $(A, \{A'\}),\ A \in N$,
3. $(A, \{A''\}),\ A \in N$,
4. $(A', \{C_i\})$,
5. $(C_i, \{C\})$,
6. $(B'', \{D_i\})$,
7. $(D_i, \{D\})$, for $q_i : AB \to CD \in P_2,\ 1 \leq i \leq m$.

Denote by N', N'' the set of symbols A', A'', respectively, $A \in N$.

Take also the hypothesis language R for σ

$$R = R_0 \cup (N \cup T)^* N' (N \cup T)^*,$$

where

$$\begin{aligned} R_0 &= (N \cup T)^* - T^* \\ &\cup (N \cup T)^* N' N'' (N \cup T)^* \cup \bigcup_{i=1}^{m} (N \cup T)^* C_i N'' (N \cup T)^* \\ &\cup \bigcup_{i=1}^{m} (N \cup T)^* C_i D_i (N \cup T)^* \cup \bigcup_{i=1}^{m} (N \cup T)^* D_i (N \cup T)^*. \end{aligned}$$

We have

$$L_{=1}(\sigma) = L(G).$$

Indeed, rules in P_1 are simulated by components in group 1 and conversely. In every moment, one component of type 2 can be used (not more, due to restrictions imposed by R). Then a component of type 3 can be used (and only one). No component of type 4 can be used before using the component of type 3. Now a component of type 5 can be used, then one of type 6 and finally one of type 7. These components must correspond to the same index i, hence to the same rule q_i in P_2, which is simulated in this way. The components of type $2-7$ cannot be used in a different order as above, due to the forms of strings in R.

The strings in R also ensure the equalities $L_{=1}(\sigma) = L_*(\sigma) = L_{\geq 1}(\sigma) = L_{\leq k}(\sigma)$, hence for all these cases we have the equality with $L(G)$.

For the derivation mode t we construct the colony σ' which is exactly as σ above, but with components of types 2, 3 replaced by

$$\begin{aligned} &2'. \quad (A, \{A', \bar{A}\}),\ A \in N, \\ &3'. \quad (\bar{A}, \{A'', A\}),\ A \in N. \end{aligned}$$

Denote again by $N', N'', \bar{N}$ the sets of symbols $A', A'', \bar{A}, A \in N$. Then the hypothesis language of σ' is

$$R' = R_0 \cup (\bar{N} \cup T)^* N' (\bar{N} \cup T)^*,$$

with R_0 as above.

The equality $L(G) = L_t(\sigma')$ can be obtained as previously (the t-mode of derivation makes necessary the use of symbols $\bar{A}$, for the case when more occurrences of some A are present, but only one can be replaced by A', A'', respectively, as requested by R'). □

Another way for increasing the power of grammar systems is considered in [15]: it is supposed that the components "speak different languages" and a transducer is necessary to intermediate them. Characterizations of context-sensitive languages are obtained in this way [14]; the same result holds true for colonies.

Definition 5. A *colony-transducer pair* is a couple (σ, g), where $\sigma = (T, (S_1, F_1), \ldots, (S_n, F_n), w)$ is a colony as above and $g = (N \cup T,\ N \cup T,\ Q, s_0, F, P)$ is a generalized sequential machine (gsm) (Q is the set of states, s_0 is the initial state, F the set of final states, P the set of translation rules of the form $sa \rightarrow xs',\ s, s' \in Q$, $a \in N \cup T,\ x \in (N \cup T)^+$). The language generated by (σ, g) in the mode f, f as above, is

$$\begin{aligned} L_f(\sigma, g) = \{x \in T^* \mid w \Longrightarrow_{i_1}^{f} w_1 \Longrightarrow g(w_1) \Longrightarrow_{i_2}^{f} w_2 \Longrightarrow g(w_2) \Longrightarrow \ldots \\ \ldots \Longrightarrow_{i_s}^{f} w_s = x,\ s \geq 1,\ 1 \leq i_j \leq n,\ 1 \leq j \leq s\}. \end{aligned}$$

(Notice that if the string is terminal, then it is no more translated.)

We denote by $\mathrm{TCOL}_n(f)$, $\mathrm{TCOL}(f)$, $n \geq 1$, f as above, the obtained families of languages.

As the transducer g can check whether the scanned string is in a given regular language R, it can play the role of a hypothesis language. Consequently,

Theorem 9. $\mathrm{TCOL}(f) = \mathrm{CS},\ f \in \{*, t, =1, \geq 1\} \cup \{\leq k \mid k \geq 1\}$.

For $f \in \{=k, \geq k \mid k \geq 2\}$ again we cannot generate strings of length strictly smaller than k. However we have

Theorem 10. $\mathrm{COL}(f) \subset \mathrm{HCOL}(f) \cap \mathrm{TCOL}(f), f \in \{=k, \geq k \mid k \geq 2\}$.

Proof. Consider the language

$$L_k = \{(a^n cb^n)^{k-1} cc^n \mid n \geq 1\}.$$

In view of Theorem 6, this language is not in $\mathrm{COL}(f)$, for f as above (we have to pump three different subwords, a^n, b^n, c^n, while in Theorem 6 only two different subwords can be pumped, on several positions, depending on f), but we have $L_k = L(\sigma)$ for

$$\sigma = (\{a, b, c\}, (S_1, \{aS_2b, cS_2, c\}), (S_2, \{S_1\}), R, S_1^k),$$

where

$$R = (a^* S_1 b^*)^{k-1} c^* S_1 \cup (a^* S_2 b^*)^{k-1} c^* S_2.$$

Clearly, the first $k-1$ occurrences of S_1 must be replaced in the first component by aS_2b and the last one by cS_2, or all of them by c, otherwise the restriction imposed by R is violated. This ensures the equality $L_k = L_f(\sigma)$ for $f \in \{=k, \geq k\}$.

As above, the language R can be replaced by a gsm which can scan only strings in R (without modifying them), hence we have also the relation $L_k \in \mathrm{TCOL}(f)$. □

As a conclusion of all these results we may state that the colonies, with various modes of derivation and having or not hypothesis languages or aided by transducers, have a considerable generative power, in spite of the small complexity of components (in fact, the components are the simplest we can imagine in Chomsky hierarchy, regular grammars generating finite languages). And, of course, we have to stress the richness of this subject from mathematical point of view, a statement already illustrated in grammar systems theory [4].

ACKNOWLEDGEMENT

Useful remarks by dr. Alexandru Mateescu on an earlier version of this paper are gratefully acknowledged.

(Received April 13, 1993.)

REFERENCES

[1] R. A. Brooks: Intelligence without representation. Artificial Intelligence *47* (1991), 139–159.

[2] R. A. Brooks: Intelligence without reason. AI Memo 1293, MIT AI Laboratory, Cambridge, Mass., 1991.

[3] E. Csuhaj-Varju and J. Dassow: On cooperating distributed grammar systems. J. Inform. Proc. Cybern. EIK *26* (1990), 49–63.

[4] E. Csuhaj-Varju, J. Dassow, J. Kelemen and Gh. Păun: Grammar Systems. Gordon and Breach, London 1994.

[5] J. Dassow: Cooperating distributed grammar systems with hypothesis languages. J. Exp. Theor. Artif. Intell. *3* (1991), 11–16.

[6] J. Dassow, J. Kelemen and Gh. Păun: On parallelism in colonies. Cybernetics and Systems *24* (1993), 37–49.

[7] J. Dassow and Gh. Păun: Regulated Rewriting in Formal Language Theory. Springer-Verlag, Berlin – Heidelberg 1989.

[8] J. Gruska: Some classifications of context-free languages. Inform. and Control *14* (1969), 152–179.

[9] J. Gruska: Descriptional complexity of context-free languages. Proc. MFCS Symp., High Tatras, 1973, 71– 84.

[10] J. Kelemen: Syntactical models of distributed cooperative systems. J. Exp. Theor. Artif. Intell. *3* (1991), 1–10.

[11] J. Kelemen and A. Kelemenová: A subsumption architecture for generative symbol systems. Cybernetics and Systems Research '92, Proc. 11th European Meeting Cybern. Syst. Res. (R. Trappl, ed.), World Scientific, Singapore, 1992, pp. 1529–1536.

[12] J. Kelemen and A. Kelemenová: A grammar-theoretic treatment of multiagent systems. Cybernetics and Systems *23* (1992), 621–633.

[13] A. Kelemenová and E. Csuhaj-Varju: Languages of colonies. 2nd Intern. Coll. Words, Language, Combinatorics, Kyoto 1992.

[14] H. C. M. Kleijn and G. Rozenberg: A study in parallel rewriting systems. Inform. and Control *44* (1980), 134–163.

[15] V. Mitrana: Pairs grammar systems – transducers. Ann. Univ. Buc., Series Matem.-Inform. *39* (1990), 73–81.

[16] A. Salomaa: Formal Languages. Academic Press, New York – London 1973.

Dr. Gheorghe Păun, Institute of Mathematics of the Romanian Academy of Sciences, PO Box 1–764, Bucuresti 70700. Romania.