数理解析研究所講究録
第 458 巻 1982 年 286-298

REPEATABLE WORDS FOR SUBSTITUTION

Taishin Nishida & Youichi Kobuchi
Department of Biophysics, Kyoto University

Introduction

Among many language defining mechanisms, sequential rewriting systems, or grammars, and parallel rewriting systems, or L systems, are the two major ways to generate the words of a language systematically. Many comparative studies of the generative powers between the parallel rewriting systems and the grammars have been investigated. It is already known that the family of languages generated by 0L systems, i.e., interactionless L systems, and that of context free languages are mutually incomparable [1,6]. In fact very simple 0L systems can generate non-context free languages, e.g., let $S=\langle\{a\},\tau,a\rangle$ where τ is a homomorphism on a^* given by $\tau(a)=a^2$, then the language generated by S is $L(S)=\{a^{2^i} \mid i\geqslant 0\}$.

Several works have been done to answer the questions why some of the parallel rewriting systems can generate such rather complicated languages, or conversely, which parallel rewriting system merely generates a context free or a rational language. Lindenmayer [4] showed a sufficient condition for a 0L system to generate a context free language. Král [3] showed a similar condition for an iterated substitution which generates a context free set. Herman and Walker [2] characterize context free languages with 0L systems through "adult" concept; i.e., a word w is adult for a 0L system if the descendants of w produced by the 0L system consist of w alone. Nishida and Kobuchi [5], and Sakarovitch, Nishida, and Kobuchi [7] introduce a recurrent word which is a generalization of an adult word; i.e., a word w is recurrent if it is a descendant of any of its descendant. Clearly an adult word is recurrent. It is shown in [7] that there is the same characterization of context free languages using iterated substitutions and the recurrent concept.

In [7] it is also shown that the set of recurrent words for a rational (resp. context free) substitution is rational (resp. context free). That is, for any parallel rewriting process, the set of recurrent words is not parallel any more. In this paper we will show that the same statement is valid for the wider set of words which contains the set of recurrent words properly. We first define repeatable word, i.e., a word w is repeatable for a substitution if it is a descendant of itself. Needless to say a recurrent word is repeatable. Then we will show that the set of repeatable words for a rational (resp. context free) substitution is rational (resp. context free). We will show our result with a similar method to that of [7]. Hence we omit the details of some technical lemmas the reader should refer to [7].

1. Preliminaries

Let Σ be a nonempty set, called alphabet, the elements of which are called letters. The finite sequences of letters, called words, together with the operation of concatenation form the free momoid generated by Σ, Σ^* ; the empty word, denoted by 1, is the identity element of Σ^*.

A subsequence of a word s is called a sparse subword of s or, for short, a subword of s. The length of a word s is, by definition, the length of the sequence s and is denoted by $|s|$; if V is any subset of Σ, $|s|_V$ denotes the number of occurrences of letters of V in s. We denote by $\mathcal{P}(\Sigma)$ the power set of Σ. The structure of monoid of Σ^* extends to the power set $\mathcal{P}(\Sigma^*)$ by $XY=\{xy \mid x\in X\ y\in Y\}$ for any subsets X and Y of Σ^*. We denote by card(Σ) the cardinality of the set Σ.

A multivalued mapping $\tau: \Sigma^* \to \Theta^*$ is a substitution if it is a homomorphism from Σ^* into $\mathcal{P}(\Theta^*)$. Thus a substitution is completely defined by the family of sets $\{\tau(a) \mid a\in\Sigma\}$ and we have $\tau(1)=1$. As any relation, a substitution $\tau: \Sigma^* \to \Theta^*$ is extended additively to $\mathcal{P}(\Sigma^*)$ by $\tau(L)=\bigcup_{f\in L}\tau(f)$ for every L in $\mathcal{P}(\Sigma^*)$.

Unless otherwise stated, we treat in this paper substitutions $\tau: \Sigma^* \to \Sigma^*$ and we call such τ a substitution on Σ^*. In this case we define for every integer n the products $\tau^{n+1}=\tau(\tau^n)$ to be those of relations ; these products are again substitutions. We shall use the following notations : $\tau^*=\bigcup_{k\geq 0}\tau^k$ and $\tau^+=\tau(\tau^*)$, where τ^0 is the identity mapping of Σ^*.

Let $u=x_1x_2\ldots x_\ell$ $x_i\in\Sigma$ and $v=s_1s_2\ldots s_\ell$ $s_i\in\Sigma^*$ be two words. The word v is said to be a descendant of u if v belongs to $\tau^n(u)$ for some positive integer n. The derivation δ from u to v is an ℓ-tuple of pairs

$\delta=((x_1,s_1),(x_2,s_2),\ldots,(x_\ell,s_\ell))$ where $s_i \in \tau^n(x_i)$ $i=1,2,\ldots,\ell$.

A substitution τ on Σ^* is said to be <u>finite</u> (resp. <u>rational</u>, <u>context free</u>) if for every a of Σ, $\tau(a)$ is a finite (resp. rational, context free) subset of Σ^*.

In the literature on L-system a pair (Σ,τ) is called a <u>0L-scheme</u> if τ is a finite substitution on Σ^*.

Let alph be the function from Σ^* into $\mathcal{P}(\Sigma)$ defined as follows : for any word f in Σ^*, alph(f) is the smallest subset S of Σ such that f is in S^*. The canonical additive extension of alph is thus a function from $\mathcal{P}(\Sigma^*)$ into $\mathcal{P}(\Sigma)$.

Let τ be a substitution on Σ^*. The <u>alphabetical projection</u> of τ, denoted by ψ_τ, is the mapping from Σ into $\mathcal{P}(\Sigma)$ defined by

$$\psi_\tau(a)=\text{alph}(\tau(a)).$$

The canonical additive extension of ψ_τ is a function from $\mathcal{P}(\Sigma)$ into itself and is again denoted by ψ_τ.

LEMMA 1.1 [7] : $\psi_\tau^2 = \psi_{\tau^2}$.

DEFINITION : A substitution τ on Σ^* is said to be <u>alphabetically stable</u> (or <u>stable</u> for short) if the following two conditions hold :

i) $\psi_\tau = \psi_\tau^2$

ii) For every a in Σ, if 1 is in $\tau^+(a)$ then 1 is in $\tau(a)$.

PROPOSITION 1.2 [7] : <u>For every substitution</u> τ <u>there exists an integer</u> r <u>such that</u> τ^r <u>is stable</u>.

2. Repeatable words

DEFINITION : Let τ be a substitution on Σ^*. A word w of Σ^* is <u>repeatable</u> for τ if it is a descendant of itself, i.e., $w \in \tau^+(w)$. We denote by $P(\tau)$ the set of repeatable words for τ:

$$P(\tau) = \{w \mid w \in \tau^+(w)\}.$$

From the above definition a repeatable word u has at least one derivation δ from u to u. If $u=x_1x_2\ldots x_\ell$ and $\delta=((x_1,s_1),(x_2,s_2),\ldots,(x_\ell,s_\ell))$ then u is factorized into $u=s_1s_2\ldots s_\ell$. We sometimes call this a factorization by δ.

Let $R(\tau)$ be the set of recurrent words for substitution on Σ^* [5,7]. That is, $R(\tau)=\{w\in\Sigma^* \mid w\in\tau^+(f) \text{ for any } f\in\tau^+(w)\}$. Then it is clear that $P(\tau)\supset R(\tau)$, and that a repeatable word is a very natural extension of that of recurrent words.

The followings are immediate consequences of the definition and have the corresponding version for recurrent words in [7], so we omit the proofs:

LEMMA 2.1 : $P(\tau)$ <u>is closed under product</u>.

LEMMA 2.2 : <u>For any positive integer</u> n, $P(\tau)=P(\tau^n)$.

3. Classification of letters

DEFINITION : Let τ be a substitution on Σ^*. A letter x of Σ is said to be <u>vital</u> for τ if 1 is not a descendant of x, i.e., if $1 \notin \tau^+(x)$. We denote by V the set of vital letters. The set of non-vital letters is denoted by N, i.e., $N=\Sigma\setminus V$, or equivalently, $N=\{x \mid 1 \in \tau^+(x)\}$.

PROPERTY 3.1 : Let $u=x_1x_2\ldots x_\ell$ $x_i \in \Sigma$ be repeatable word. Let $\delta=((x_1,s_1),(x_2,s_2),\ldots,(x_\ell,s_\ell))$ be a derivation from u to u. Then

1. If x_i is non-vital then $|s_i|_V=0$.
2. If x_i is vital then x_i is the only vital letter contained in s_i.

DEFINITION : Let τ be a substitution on Σ^*. A letter x in Σ is said to be cyclic if there exist two words s and t in N^* such that sxt is in $\tau^+(x)$. The set of cyclic letters is denoted by C.

DEFINITION : Let τ be a substitution on Σ^*. Let V be the set of vital letters for τ. τ is said to be vitality preserving if for any x in Σ and any u in $\tau(x)$, $|u|_V=|x|_V$.

Let τ be a substitution on Σ^*. Let V, N, and C the set of vital, non-vital, and cyclic letters for τ, respectively. Consider the mapping τ' on $(N\cup C)^*$ defined by

$$\begin{cases} \tau'(x)=\tau(x)\setminus\Sigma^*V\Sigma^*V\Sigma^* & \text{if } x\in C\cap V, \\ \tau'(x)=\tau(x)\setminus\Sigma^*V\Sigma^* & \text{if } x\in N. \end{cases}$$

It is easily seen that τ' is a well defined substitution on

(NUC)* and τ' is vitality preserving. We call τ' the vitality preserving substitution of τ. The following is an immediate consequence of Property 3.1.

PROPERTY 3.2 : $P(\tau)=P(\tau')$.

That is, we may also assume, without loss of generality, that τ is vitality preserving to compute the set of repeatable words.

LEMMA 3.3 [7] : Let τ be stable substitution on Σ^*. For any f in C* there exists a word s such that f is a subword of s and s is in $\tau^k(f)$ for every positive integer k.

COROLLARY 3.4 : Let τ be a stable and vitality preserving substitution on Σ^*. For any x in C and any positive integer k

$$\tau^k(x) \subset \tau^{k+1}(x).$$

4. Letter position function

In this section we define a letter position function, which turns out to be very useful in the following discussion. We characterize the letters which appear in a repeatable word using the letter position function.

DEFINITION : Let $u=x_1x_2\ldots x_\ell$ $x_i \in \Sigma$ be a repeatable word for a substitution τ on Σ^*. Let $\delta=((x_1,s_1),(x_2,s_2),\ldots,(x_\ell,s_\ell))$ be the derivation from u to u. The letter position function $\alpha : \{1,2,\ldots,\ell\} \to \{1,2,\ldots,\ell\}$ for δ is given by $\alpha(i)=j$ where x_i is contained in s_j.

Informally, α indicates the ancestor of each letter of u in the derivation δ. It is very important, although obvious by the definition, that α is a non-decreasing function on the integer interval $[1,\ell]$. Then we have

PROPERTY 4.1 : Let α be a letter position function. α has at least one fixed point, i.e., there exists an i in $\{1,2,\ldots,\ell\}$ such that $\alpha(i)=i$.

DEFINITION : Let $u=x_1x_2\ldots x_\ell$ be a repeatable word. Let α be the letter position function for a derivation δ from u to u. Then x_i is said to be repeatable in u if i is a fixed point of α.

LEMMA 4.2 : Let $u=x_1x_2\ldots x_\ell$ be a repeatable word. If x_i is vital then x_i is repeatable in u.

Proof : Let α be the letter position function for a derivation δ from u to u. Let x_i be a vital letter and let $\alpha(i)=j$. From Property 3.1.1 x_j must be vital. And then, from Property 3.1.2, we have i=j. □

There is a very close relation between cyclic letter and repeatable letter in a repeatable word. Indeed we have

PROPOSITION 4.3 : If x_i is repeatable in u then x_i is in C.

Proof : Let s_i be the segment produced by x_i in a derivation from u to u. Then, from Lemma 4.2, x_i is the only possible vital letter in s_i, even if s_i contains any vital letters. Therefore, for some words f and g in N^*, we have $s_i = fx_ig$ and s_i is in $\tau^+(x_i)$. □

5. Characterization of repeatable words

DEFINITION : A repeatable word u for τ is said to be elementary if $u \neq 1$ and any factorization $u = u_1u_2$ for repeatable words u_1 and u_2 implies $u_1 = 1$ or $u_2 = 1$. We denote by $P_1(\tau)$ the set of elementary repeatable words for τ.

PROPERTY 5.1 : $P(\tau) = (P_1(\tau))^*$.

Proof : $P(\tau) \supset (P_1(\tau))^*$ is obvious by Lemma 2.1. $P(\tau) \subset (P_1(\tau))^*$ directly follows from the definition. □

LEMMA 5.2 : Let $u = x_1x_2 \ldots x_\ell$ be a word in $P_1(\tau)$. There exists one and only one i such that x_i is repeatable in u.

Proof : Assume there are exactly two cyclic letters x_i, x_j $(i<j)$ for a letter position function α. (We can show the lemma similarly in case there are more than two cyclic letters.)

As i and j are the fixed points of α, there exists a positive integer p and an integer k $i \leq k < j$ such that

$$\alpha^p(n) = \begin{cases} i & \text{if } 1 \leq n \leq k \\ j & \text{if } k+1 \leq n \leq \ell. \end{cases}$$

That is, $t_1 = x_1 x_2 \ldots x_k$ is in $\tau^+(x_i)$ and $t_2 = x_{k+1} \ldots x_\ell$ is in $\tau^+(x_j)$. Since $u = t_1 t_2$ and u is in $\tau^+(u)$, t_1 is in $\tau^+(t_1)$ and t_2 is in $\tau^+(t_2)$. Thus t_1 and t_2 are repeatable. This contradicts the fact that u is elementary. □

The proof of the above lemma also shows

LEMMA 5.3 : Let $u = x_1 x_2 \ldots x_\ell$ be a word in $P_1(\tau)$ and x_i be repeatable in u. Then u is in $\tau^+(x_i)$.

PROPOSITION 5.4 : Let τ be a stable substitution on Σ^*. Let C and N be the set of cyclic and non-vital letters for τ, respectively. Then

$$P_1(\tau) \subset \bigcup_{x \in C} (\tau^+(x) \cap N^* x N^*) \subset P(\tau).$$

Proof : The left side inclusion is a corollary of the above Lemmas. Let u be a word in $\tau^+(x) \cap N^* x N^*$ for some $x \in C$. Then u is written as $u = sxt$ for some st in N^*. Since τ is stable, 1 is in $\tau(st)$, and hence u is in $\tau^+(sxt) = \tau^+(u)$. □

COROLLARY 5.5 : $P(\tau) = (\bigcup_{x \in C} (\tau^+(x) \cap N^* x N^*))^*$.

Next we will give an effective finitary description of

$\bigcup_{x\in C}(\tau^{+}(x)\cap N^{*}xN^{*})$. We need some additional definitions. Throughout the remaining of this section let τ be a stable and vitality preserving substitution.

For any x in C let C_x be the set defined by

$$C_x=\{y \mid \exists y_0, y_1, \dots, y_n \in C \text{ such that } y_0=y_n=x \ \exists k \ y_k=y \text{ and for any } i \ 0<i\leq n \ \exists s_i, t_i\in N^* \ s_i y_i t_i \in \tau(y_{i-1})\}.$$

Let D_x, E_x, G_x, and H_x be the sets defined by

$$E_x=\text{alph}(\{g \mid \exists f\in N^*, \exists y,z\in C_x \ fyg\in\tau(z)\})$$

$$H_x=\text{alph}(\{f \mid \exists g\in N^*, \exists y,z\in C_x \ fyg\in\tau(z)\})$$

$$D_x=\psi_\tau(E_x)\cap N\cap C$$

$$G_x=\psi_\tau(H_x)\cap N\cap C.$$

PROPOSITION 5.6 : <u>There exists an integer</u> n <u>such that for every</u> x <u>in</u> C

$$\tau^{+}(x)\cap N^{*}xN^{*}=\tau^{n}(G_x^{*}xD_x^{*})\cap N^{*}xN^{*}.$$

In order to prove this Proposition we establish two lemmas, which can be shown quite similarly as in the proof of the corresponding results in [7].

LEMMA 5.7 : <u>Let</u> x <u>be a cyclic letter.</u> <u>For any word</u> fxg <u>in</u> $G_x^{*}xD_x^{*}$ <u>there exists a word</u> sxt <u>in</u> $\tau^{+}(x)$ <u>such that</u> fxg <u>is a subword of</u> sxt.

LEMMA 5.8 : <u>There exists an integer</u> k <u>with the following property</u> : <u>Let</u> x <u>be any cyclic letter,</u> sxt <u>any word in</u> $\tau^{*}(x)\cap N^{*}xN^{*}$, and c any letter of s (resp. of t); if c is not in $\psi_\tau(G_\chi)$ (resp. in $\psi_\tau(D_\chi)$) there exists a word uxv in $\tau^k(x)\cap N^*xN^*$ such that $sxt=s'uxvt'$ and c does not occur in s' (resp. in t').

Proof of Proposition 5.6 :

We first prove that

$$\tau^+(G_\chi^*x\, D_\chi^*)\subset \tau^+(x)$$

for any cyclic letter x. Let fxg be any word of $G_\chi^*xD_\chi^*$. From Lemma 5.7 there exist an integer k and a word sxt such that fxg is a subword of sxt and sxt is in $\tau^k(x)$. Now any descendant w of fxg is a descendant of sxt and thus of x.

Conversely let w be in $\tau^+(x)\cap N^*xN^*$: there exist z in C_χ, and s,t in N^* such that w is in $\tau(szt)$, i.e., there exist two subwords $f=a_1a_2\ldots a_p$ and $g=b_1b_2\ldots b_q$ of s and t respectively and a factorization $w=s_1s_2\ldots s_puyvt_1t_2\ldots t_q$ such that s_i is in $\tau(a_i)$, t_j in $\tau(b_j)$ for every i and j, and uyv is in $\tau(z)$. If all the a_i's and the b_j's are in $\psi_\tau(G_\chi)$ and $\psi_\tau(D_\chi)$ respectively w is in $\tau^2(G_\chi^*xD_\chi^*)$. If it is not the case it follows from Lemma 5.8 that there exists a factorization $szt=s'u'zv't'$ such that all the a_i occurring in s' are in $\psi_\tau(G_\chi)$, all the b_j occurring in t' are in $\psi_\tau(D_\chi)$ and uzv is in $\tau^k(z)$. Then w is in $\tau^{k+1}(G_\chi^*xD_\chi^*)$. □

THEOREM 5.9 : The set of repeatable words for a rational (resp.context free) substitution is a rational (resp.context free) set.

Proof : The families of rational languages and context free languages are both closed under union, intersection with rational set, Kleene closure, and substitution. Then the theorem follows immedeately from Corollary 5.5 and Proposition 5.6. □

REFERENCES

[1] Herman, G.T. & Rozenberg, G. : Developmental systems and languages, North-Holland, Amsterdam, 1974.

[2] Herman, G.T. & Walker, A. : Context free languages in biological systems, Inter. J. of Computer Math., 4, 369-391, 1975.

[3] Král, J. : A modification of a substitution theorem and some necessary and sufficient conditions for sets to be context free, Math. Systems Teory, 4, 129-239, 1970.

[4] Lindenmayer, A. : Developmental systems without cellular interactions, their languages and grammars, J. Theor. Biol., 30, 455-484, 1971.

[5] Nishida, T. & Kobuchi, Y. : Recurrent strings in a 0L language, 数理解析研究所講究録, 421, 121-133, 1981.

[6] Rozenberg, G. & Salomaa, A. : The mathematical theory of L systems, Academic Press, New York, 1980.

[7] Sakarovitch, J., Nishida, T. & Kobuchi, Y. : Recurrent words for substitution, Technical report of Research Institute for Mathematical Sciences Kyoto University, RIMS-396, 1982.