Marcus Contextual Languages Consisting of Primitive words [1]

Pál DÖMÖSI
Institute of Informatics, Debrecen University
Debrecen, Egyetem tér 1., H-4032, Hungary
e-mail: domosi@math.klte.hu

and

Masami ITO
Kyoto Sangyo University
Faculty of Science, Kyoto 603-8555 Japan
e-mail: ito@ksuvx0.kyoto-su.ac.jp

and

Solomon MARCUS
Istitute of Mathematics of the Romanian Academy of Sciences
Bucharest, Romania
e-mail: smarcus@stoilow.imar.ro

Abstract

In this paper we prove that the language of all primitive words over a non-trivial alphabet can be generated by certain types of Marcus contextual grammars. Some open problems are also discussed.

Keywords: Formal languages and automata, combinatorics of words and languages.

[1] The first author of this work was supported by grant from Japan Society for Promotion of Science (No.) and Xerox Foundation UAC grant (1478-2004), U.S.A. .

1 Introduction

Marcus contextual grammars play an important role in theoretical computer science. (See, for example, Gh. Păun [13].) It is a well-known conjecture of P. DÖMÖSI, S. HORVÁTH, M. ITO [3] that the language Q of all primitive words over a nontrivial alphabet (having at least two letters) is not context-free. This conjecture is not proved or improved so far. It is a natural and interesting question whether or not Q can be generated by some other well-known types of grammars. Now we consider three types of Marcus contextual grammars from this point of view.

2 Preliminaries

A *word* (over Σ) is a finite sequence of elements of some finite non-empty set Σ. We call the set Σ an *alphabet,* the elements of Σ *letters.* If u and v are words over an alphabet Σ, then their *catenation* uv is also a word over Σ. Especially, for every word u over Σ, $u\lambda = \lambda u = u$, where λ denotes the *empty word.* Given a word u, we define $u^0 = \lambda,\ u^n = u^{n-1}u,\ n > 0,\ u^* = \{u^n : n \geq 0\}$ and $u^+ = u^* \setminus \{\lambda\}$.

The *length* $|w|$ of a word w is the number of letters in w, where each letter is counted as many times as it occurs. Thus $|\lambda| = 0$. By the *free monoid* Σ^* *generated by* Σ we mean the set of all words (including the *empty word* λ) having catenation as multiplication. We set $\Sigma^+ = \Sigma^* \setminus \{\lambda\}$, where the subsemigroup Σ^+ of Σ^* is said to be the *free semigroup generated by* Σ. Subsets of Σ^* are referred to as *languages* over Σ.

A *primitive word* (over Σ, or actually over an arbitrary alphabet) is a nonempty word not of the form w^m for any nonempty word w and integer $m \geq 2$. The set of all primitive words over Σ will be denoted by $Q(\Sigma)$, or simply by Q if Σ is understood. Q has received special interest: Q and $\Sigma^+ \setminus Q$ play an important role in the algebraic theory of codes and formal languages (see M. LOTHAIRE [10] and H. J. SHYR [14]).

Denote by $|H|$ the *cardinality* of H for every set H.

The next statement is from H.J. SHYR and G. THIERRIN[14, 16].

Theorem 2.1 *Let $i \geq 1$ and $uv \in \{p^i : p \in Q\}$. Then $vu \in \{p^i : p \in Q\}$, too. In other words, the sets $\{p^i : p \in Q\}$ $(i \geq 1)$ are closed under cyclic permutations of words.* □

We shall use the following result of H.J. SHYR [14, 15].

Theorem 2.2 *Let $f, g \in Q, f \neq g$ and $n \geq 1$. If $fg^n \notin Q$ then $fg^{n+k} \in Q$ for all $k \geq 2$.* □

Let $u \neq \lambda$ and let f be a primitive word with an integer $k \geq 1$ having $u = f^k$. We let $\sqrt{u} = f$ and call f the *primitive root* of the word u. R.C. LYNDON, M.P. SCHÜTZENBERGER proved in [11] the uniqueness of primitive root as follows. (See also H.J. SHYR [14].

Theorem 2.3 *If $u \neq \lambda$, then there exists a unique primitive word f and a unique integer $k \geq 1$ such that $u = f^k$.* □

The next statement is also from R.C. LYNDON, M.P. SCHÜTZENBERGER [11].

Theorem 2.4 *Let $f, g \in Q, f \neq g$. Then $f^m g^n \in Q$ for all $m \geq 2, n \geq 2$.* □

The following result by N.J FINE and H.S. WILF [4, 6] will also be applied. (For a version of this statement see also H.J. SHYR [14].)

Theorem 2.5 *Let u and v be nonempty words, and, $p, q \geq 0$ integers. If u^p and v^q contain a common prefix or suffix of length $|u| + |v| - g.c.d.(|u|, |v|)$ (where $g.c.d.(|u|, |v|)$ denotes the greatest common divisor of $|u|$ and $|v|$) then $u = w^m$ and $v = w^n$, for some word w and positive integers m, n.* □

Next we prove

Lemma 2.6 *Let Σ be a nontrivial alphabet (with $|\Sigma| \geq 2$). If $w, wa \notin Q$, where $w \in \Sigma^+$ and $a \in \Sigma$, then $w \in a^+$.*

Proof: Since $w, wa \notin Q$, there exist $p, q \in Q$ and $m, n \geq 2$ such that $w = p^m$ and $wa = q^n$. Hence $|p| = |w|/m \leq |w|/2$ and $|q| = (|w| + 1)/n \leq (|w| + 1)/2$. Using Theorem 2.5, these inequalities imply that $|p| + |q| - g.c.d(|p|, |q|) < |w|$ and $p = q$. From the fact that $1 = |wa| - |w| = (m - n)|p|$, it follows that $|p| = 1$ and $p = a$. This completes the proof of the lemma. □

The next result is from H.J. SHYR, S.S. YU [17].

Theorem 2.7 *Let $p, q \in Q, p \neq q$. Then $|p^+q^+ \cap \Sigma^+ \setminus Q| \leq 1$.* □

We shall use the following statement of G. BORWEIN [14].

Theorem 2.8 *Let $u \in \Sigma^+, u \neq a^n, a \in \Sigma$. Then one of ua, u must be primitive.* □

A *(Marcus) contextual grammar with choiche* is a structure $G = (V, A, C, \varphi)$, where V is an alphabet, A is a finite language over V, C is a finite subset of $V^* \times V^*$, and $\varphi : V^* \to 2^C$. If $\varphi(x) = C$ holds for every $x \in X^*$ then we say that G is a (Marcus) contextual grammar without choiche and then we omit φ sometimes.

We define two relations on V^* as usual: for any $x \in V^*$, we write
$x \Rightarrow_{ex} y$ if and only if $y = uxv$, for a context (u, v) in $\varphi(x)$,
$x \Rightarrow_{int} y$ if and only if $x = x_1x_2x_3, y = x_1ux_2vx_3$ for any $(u, v) \in \varphi(x_2)$.
Denote $\overset{*}{\Rightarrow}_{ex}$, $\overset{*}{\Rightarrow}_{in}$ the reflexive and transitive closure of these relations and let $L_\alpha(G) = \{x \in V^* : w \overset{*}{\Rightarrow}_\alpha x, w \in A\}$ for $\alpha \in \{ex, in\}$. Then $L_{ex}(G)$ is the *(Marcus) external contextual language (with or without choiche) generated by* G, and similarly, $L_{in}(G)$ is the *(Marcus) internal contextual language (with or without choiche) generated by* G. Now let $G = (V, A, \varphi)$, where V is an alphabet, A is a finite language over V, C is a finite subset of $V^* \times V^*$, and $\varphi : V^* \times V^* \times V^* \to 2^C$.[2]

Define the relation $\Rightarrow$ on V^* such that $x \Rightarrow y$ for some $x, y \in V^*$ if and only if $x = x_1x_2x_3, y = x_1ux_2vx_3, x_1, x_2, x_3 \in V^*, (u, v) \in \varphi(x_1, x_2, x_3)$. Moreover, let $\overset{*}{\Rightarrow}$ denote the reflexive and transitive closure of $\Rightarrow$. Thus $L(G)$ is defined to be a *(Marcus) total contextual grammar (with or without choiche) generated by* G. If $\varphi(x_1, x_2, x_3) = C$ holds for every $x_1, x_2, x_3 \in X^*$ then we say that G is a (Marcus) total contextual grammar without choiche and sometimes we omit φ having this property.

Theorem 2.9 *The language Q of all primitive words over a nontrivial alphabet V is an external contextual language (with choiche).*

Proof: Define $G = (V, A, C, \varphi)$ in the following manner. Let $A = V$, $C = \{(\emptyset, a) : a \in V^+, |a| \in \{1, 2\}\}$, moreover, let for every $w \in V^*$, $\varphi(w) = \{(\emptyset, x) : (\emptyset, x) \in C, wx \in Q\}$. Clearly, then $L_{ex}(G) \subseteq Q$. Then it is enough to show that $w \in L_{ex}(G)$ holds for every $w \in Q$. By the definition of A this holds if $w \in A$, i.e. $|w| = 1$. If $|w| = 2$ and $w \in Q$ then $w = ab$ holds for some $a, b \in V, a \neq b$. Obviously, then $a \in A$ and $(\emptyset, b) \in \varphi(a)$. Therefore, $a \Rightarrow_{ex} ab = w \in L_{ex}(G)$. Now let $n \geq 2$ be an arbitrary positive integer such that $w \in L_{ex}(G)$ whenever $w \in Q$ and $|w| \leq n$. Consider an arbitrary $z \in Q$ with $|z| = n + 1$. Prove $z \in L_{ex}(G)$. Then $z = ub$ and $z = vab$ for some $a, b \in V, u, v \in V^*, |u| = n, |v| = n - 1, u = va$. By Theorem 2.8, one of va, v must be primitive. If $va = u$ is primitive, then $(\emptyset, b) \in \varphi(z)$ leading to $u \Rightarrow_{ex} ub = z \in L_{ex}(G)$. Similarly, if v is primitive, then $(\emptyset, ab) \in \varphi(v)$ leading to $v \Rightarrow_{ex} vab = z \in L_{ex}(G)$. This completes the proof. □

It is well-known (see, for example, [13], that the total contextual grammars are a generalization of both the internal and the external ones. Thus, by the above statement, language Q of all primitive words over a nontrivial alphabet V

[2]Observe that the definition of φ is not the same as before.

is a total contextual language (with choiche). Regarding the internal contextual grammar, we have the following statement.

Theorem 2.10 *The language Q of all primitive words over a nontrivial alphabet V is not an internal contextual language (with choiche).*

Proof: Suppose that, contrary of our assumptions, there exists a $G = (V, A, C, \varphi)$ with $Q = L_{in}(G)$. Obviously, then we may assume $(\lambda, \lambda) \notin C$. Consider a pair $x, y \in L_{in}(G), (u, v) \in \varphi(w)$ such that $x = x_1x_2x_3, y = x_1ux_2vx_3, (u, v) \in \varphi(x_2)$. Then $x \Rightarrow_{in} y \in Q$ such that $x \neq \lambda$ (because $\lambda \notin Q$.)

Suppose $x_2 = \lambda$. Then we may assume that for every $z \in Q, uvz \in Q$. By Theorem 2.4 and Theorem 2.1, we have $x(uv)^2x \in Q$ for every $x \in Q, x \neq \sqrt{uv}$. Therefore, $x(uv)^2x \Rightarrow_{in} uvx(uv)^2x \Rightarrow ((uv)^2x)^2 \in Q$, a contradiction.

Suppose $x_2 \notin \lambda$. Let $y_1 \in Q$ be arbitrary such that $y_1x_2x_3 \in Q$ and $\sqrt{y_1x_2} \neq \sqrt{y_1ux_2v}$.[3] If $y_1x_2y_1ux_2v \in Q$ then $(u, v) \in \varphi(x_2)$ implies $y_1x_2y_1ux_2v \overset{*}{\Rightarrow}_{in} (y_1ux_2v)^2 \in Q$, a contradiction. Thus $y_1x_2y_1ux_2v \notin Q$. Prove that for an appropriate $k \geq 1$, $y_1x_2(y_1ux_2v)^n \in Q, n \geq k$. Indeed, because of Theorem 2.4 and $\sqrt{y_1x_2} \neq \sqrt{y_1ux_2v}$, we obtain $y_1x_2(y_1ux_2v)^n \in Q, n \geq 2$ if $y_1x_2 \notin Q$. Otherwise, because of Theorem 2.2, $y_1x_2(y_1ux_2v)^n \in Q, n \geq 3$. But then there exists a $k \geq 1$, having $y_1x_2(y_1ux_2v)^n \Rightarrow_{in} (y_1ux_2v)^{n+1} \in Q, n \geq k$, a contradiction. This ends the proof. □

We have the following

Corollary 2.11 *The language Q of all primitive words over a nontrivial alphabet V is not an internal contextual language without choiche.* □

Theorem 2.12 *The language Q of all primitive words over a nontrivial alphabet V is not an external contextual language without choiche.*

Proof: Assume $G = (V, A, C)$ with $Q = L_{ex}(G)$. Let $(u, v) \in C$ such that $(u, v) \neq (\lambda, \lambda)$. Then for every $z \in Q$, we obtain $z \Rightarrow_{ex} uzv \in Q$ with $z \neq uzv$. Observe that $zuzv \in Q$ is impossible because of $zuzv \Rightarrow_{ex} (uzv)^2 \notin Q$. But then, applying Theorem 2.2 with $f = z, g = uzv$, we get $z(uzv)^n \in Q$ for all $n \geq 3$. But then $z(uzv)^n \Rightarrow_{ex} (uzv)^{n+1} \in Q$, a contradiction. □

Observe that we may apply the main idea of the above proof for total contextual grammars without choice such that we consider the derivations $z \Rightarrow_{ex} \lambda uzv\lambda = uzv$, $zuzv \Rightarrow_{ex} \lambda uzv\lambda uzv = (uzv)^2$ and $z(uzv)^n \Rightarrow_{ex} \lambda uzv\lambda(uzv)^n = (uzv)^{n+1}$ instead of $z \Rightarrow_{ex} uzv$, $zuzv \Rightarrow_{ex} (uzv)^2$ and $z(uzv)^n \Rightarrow_{ex} (uzv)^{n+1}$. Thus we get as follows.

[3]For example, if $y_1 = a^rb$ for some $a, b \in V, a \neq b$ such that $r > |ux_2v|$ then y_1 has this property.

Theorem 2.13 *The language Q of all primitive words over a nontrivial alphabet V is not a total contextual language without choiche.* □

Remark 2.14 *It can be easily shown that an external Marcus contextual language (without choice) is a linear context-free language. Moreover, it is shown in [7] that Q is not a linear context-free language. Therefore, we can also get Theorem 2.12 as a direct consequence of this result in [7].*

Theorem 2.15 *The language Q is an external Marcus contextual langauge (with choice).*

Proof: Let $G = (V, A, C, \varphi)$ be an external Marcus contextual grammar with choice. Notice that the proposition holds true for $|V| = 1$. Hence we assume $|V| \geq 2$. By the definition of the grammar G, it is obvious that $L_{ext}(G) \subseteq Q$. Now we prove that $Q \subseteq L_{ext}(G)$ by induction. First, we have $(V \cup V^2) \cap Q \subseteq L_{ext}(G)$. Now, assume that $(V \cup V^2 \cup \cdots \cup V^n) \cap Q \subseteq L_{ext}(G)$ for some $n \geq 2$. Let $u \in V^{n+1} \cap Q$ and let $u = wab$ where $a, b \in V^+$.

Case 1. $w \in Q$ or $wa \in Q$. In this case, by induction hypothesis, $w \in L_{ext}(G)$ or $wa \in L_{ext}(G)$. Then $w \Rightarrow_{ext} wab$ or $wa \Rightarrow_{ext} wab$, i.e. $u \in L_{ext}(G)$.

Case 2. $w \notin Q$ and $wa \notin Q$. In this case, by Lemma 1, we have $u = a^n b$. Since $u \in Q$, $a \neq b$. By the configuration: $b \Rightarrow_{ext} ab \Rightarrow_{ext} a^2 b \Rightarrow_{ext} \cdots \Rightarrow_{ext} a^n b$, we have $u \in L_{ext}((G)$. Thus $Q = L_{ext}(G)$. This means that Q is an external Marcus contextual langauge (with choice). □

3 Generalizations, open problems

Taking into account that "primitive" is very general, being the negation of a very restrictive property, "periodicity", we propose to consider "strongly primitive" as a particular form of "primitive", obtained by negation of "quasiperiodic", an extension of "periodic" as follows. The word w of length n, over the alphabet Σ of cardinal k, is quasiperiodic, with quasiperiod y, if y is a factor (i.e., subword) of w of length $|y| = m$ strictly smaller than n and each position in w falls within an occurrence of y. We say that w is n/m-quasiperiodic. For example, $w = ababa$ is 5/3-quasiperiodic, with $y = aba$ as quasiperiod. This example shows that "quasiperiodic" is an effective extension of "periodic" (excepting the case $k = 1$) Obviously, any periodic word is quasiperiodic. Now, define a strongly primitive word by the property to not be quasiperiodic. Two natural questions appear: What is the position of the language L of strongly primitive words in the Chomsky hierarchy ? What is the position of L in respect to various types of contextual languages ? One can define n/m-strongly primitive words and speculate on this infinite typology of strongly primitive words. Things may

depend sometimes on the cardinal k of Σ. Since quasiperiodicity occurs in the study of DNA sequences (see S. CARLIN, M. MORRIS, G. GHANDOUR, M. Y. LEUNG [1]), as well as in musical composition (see T. CRAWFORD, C. S. ILIOPOULOS, R. RAMAN [2]), strong primitivity may have motivations of similar nature.

We proved that the language Q of all primitive words over a nontrivial alphabet is an external contextual language with choice. (But it is an open problem whether Q belongs to context free languages too.) It is a further challenge of research to try to check what happens when instead of primitive words we consider strongly primitive words ; as we have already defined before, a finite word w over the alphabet Σ is strongly primitive if it is not quasiperiodic, i.e., there exists no strict subword u of w such that every position in w is included in an occurrence of u in w. For instance, we remarked before that $w = ababa$ is quasiperiodic, because every position in w is included in an occurrence of $u = aba$, so w is not strongly primitive. On the other hand, $w = abaab$ is not quasiperiodic, so it is quasiprimitive. Obviously, any periodic word is quasiperiodic, but the converse is not true.Every strongly primitive word is primitive, but the converse is not true. See also S. MARCUS [12] and F. LEVÉ, G. RICHOMME [9].

Extend primitivity from words to languages. Define an infinite language L on the ordered alphabet Σ to be periodic, quasiperiodic or almost periodic if the infinite word on Σ obtained by concatenation of the words in L according to their increasing length (and in their lexicographic order when they are of the same length) is periodic, quasiperiodic resp. almost periodic. We recall that an infinite word w is almost periodic if any factor of w occurs infinitely many times as factor of w and the gaps between its consecutive occurrences are bounded. Now let us say that L is primitive if it is not periodic; L is quasiprimitive if it is not quasiperiodic; L is almost primitive if it is not almost periodic. Now we can try to check what properties of primitive (quasiprimitive) words can be transfered to primitive (quasiprimitive) languages. What about almost primitive languages?

The concept of a primitive word can be extended to languages in the following way too: A language L on the alphabet Σ is primitive if all its words are primitive (but it is not obligatory for L to include all primitive words on Σ). A language L on Σ is essentially primitive if it is primitive, but there exist infinitely many primitive words on Σ which are missing from L. What about the position of primitive languages and of essentially primitive languages in the Chomsky hierarchy ? What about their position in respect to various types of contextual languages ? Are they distributed in all classes of Chomsky hierarchy ? in what types of contextual languages ? Obviously, the non-trivial case corresponds to L infinite.

References

[1] S. CARLIN, M. MORRIS, G. GHANDOUR, M. Y. LEUNG, Efficient algorithms for molecular sequences analysis. Proc. Natl. Acad. Sci., USA, **85**(1988),841-845.

[2] T. CRAWFORD, C. S. ILIOPOULOS, R. RAMAN, String matching techniques for musical similarity and melodic recognition. *Computers and Musicology,* **11**(1998), 72-100.

[3] P. DÖMÖSI, S. HORVÁTH, M. ITO, On the connection between formal languages and primitive words. In: Proc. First Session on Scientific Communication, Univ. of Oradea, Oradea, Romania, 6-8 June, 1991, Analele Univ. of Oradea, Fasc. Mat., 1991, 59–67.

[4] N.J. FINE, H.S. WILF, Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society,* **16** (1965) 109-114.

[5] S. GINSBURG, The Mathematical Theory of Context-Free Languages. *McGraw - Hill,* 1966.

[6] M. HARRISON, Introduction to Formal Language Theory. *Addison-Wesley, Reading, MA,* 1978.

[7] S. HORVÁTH, Strong interchangeability and nonlinearity of primitive words. In : Nijholt, A., Scollo, G., Steetskamp, R. eds., Proc. Works. AMiLP'95 (Algebraic Methods in Language Processing, 1995) Univ. of Twente, Enschede, the Netherlands, 6-8 Dec., 1995, Univ. Twente, Enschede, 1995, 173-178.

[8] L. ILIE, On a conjecture about slender context-free languages. *Theoret. Comput. Sci.,* **132** (1994) 427-434.

[9] F. LEVÉ, G. RICHMONDE, Quasiperiodic infinite words: some answers. *Bulletin of the European Association for Theoretical Computer Science (EATCS),* **84**(2004), 128-138.

[10] M. LOTHAIRE, Combinatorics on Words. *Addison-Wesley,* 1983.

[11] R.C. LYNDON, M.P. SCHÜTZENBERGER, The equation $a^M = b^N c^P$ in a free group. *Michigan Math. J.* **9** (1962), 289–298.

[12] S. MARCUS, Quasiperiodic infinite words. *Bulletin of the European Association for Theoretical Computer Science (EATCS),* **82**(2004), 170-174.

[13] Gh. PĂUN, Marcus Contextual Grammars. *Kluwer,* Dordrecht, Boston, London, 1997.

[14] H.J. SHYR, Free Monoids and Languages. *Ho Min Book Company, Taiwan,* 1991.

[15] H.J. SHYR, Disjunctive Languages on a Free Monoid. *Information and Control,* **64** (1977), 123–129.

[16] H.J. SHYR, G. THIERRIN, Disjunctive languages and codes. *FCT'77, LNCS* **56**, *Springer-Verlag* (1977), 171–176.

[17] H.J. SHYR, S.S. YU, Non-primitive words in the language p^+q^+. *Soochow J.Math.* **20** (1994), 535–546.