A Small Hierarchy of Languages Consisting of Non-Primitive Words[1]

PÁL DÖMÖSI
Institute of Informatics, Debrecen University
Debrecen, Egyetem tér 1, H-4032, Hungary
e-mail: domosi@inf.unideb.hu

and

GÉZA HORVÁTH
Institute of Informatics, Debrecen University
Debrecen, Egyetem tér 1, H-4032, Hungary
e-mail: geza@inf.unideb.hu

and

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

In memory of our late good friend, Professor Dr. Jürgen Duske

ABSTRACT

Context-free languages consisting of non-primitive words have been characterized by M. Ito and M. Katsura in 1988. In this paper we show that the same type of characterization can be given for linear, respectively, regular languages consisting of non-primitive words. The observation completes our knowledge on the structure of different languages of non-primitive words.

1. Introduction

A word is said to be *primitive* if it is not a repetition of another word. In other words, a word p is primitive if for any word w and $i \geq 1$, $p = w^i$ implies $i = 1$. Otherwise we speak about a *non-primitive word.* Thus, a word is called non-primitive if it is a repetition of another word. The study of relationships between the free semigroup X^+ generated by an alphabet X and the language of all primitive words Q over X has received special interest in theoretical computer science. In this paper we investigate properties of languages consisting of non-primitive words in relation with the Chomsky-hierarchy. First we establish that $L \cap Q, L \cap Q^{(i)}, i > 1, L \cap (\cup_{i>1} Q^{(i)})$ are context-sensitive languages whenever L is context-sensitive. Thus a characterization of context-sensitive languages consisting of non-primitive words has no special interest. Context-free languages consisting of non-primitive words are characterized by M. Ito and M. Katsura [1]. In this paper we also characterize regular and linear languages consisting of non-primitive words.

[1]This work was partly supported by grants of the "Automata & Formal Languages" project of the Hungarian Academy of Sciences and the Japanese Society for Promotion of Science (No 15), and the Hungarian National Foundation for Scientific Research (OTKA T030140).

2. Preliminaries

Our notions and notation concerning formal languages are standard (see, e.g., [3],[4]). Let X be a finite alphabet with more that one letter and let X^* be the free monoid generated by X. By λ we denote the identity of X^*. λ is often called *empty word.* We put $X^+ = X^* \setminus \{\lambda\}$. Thus X^+ denotes the free semigroup generated by X. An element $u \in X^*$ is called a *word* over X and u is also called a *non-empty word* if $u \in X^+$. The *length* of a word u is denoted by $|u|$. Thus $|\lambda| = 0$. Any subset L of X^* is called a *language* L over X. If there is no danger of confusion then we do not distinguish the element of a singleton set from the singleton set itself. Therefore, for instance, $\{a\}\{b\}^*\{c\}$ can be expressed as ab^*c and we write $L \setminus \lambda$ for $L \setminus \{\lambda\}$. A word u over X is said to be *primitive* if $u = v^n, v \in X^*$ implies $n = 1$. Thus the empty word is non-primitive. Throughout this paper, the set of all primitive words over X is denoted by Q. In addition, we put $Q^{(i)} = \{q^i \mid q \in Q\}$ for every $i \geq 0$. Then $Q^{(0)} = \lambda$ by definition. Note that any word $u \in X^+$ can be uniquely expressed as $u = p^n, p \in Q$ and $n \geq 1$. (See Lyndon, R. C. and Schützenberger, M. P. [2].) Hence $X^* = \cup_{i \geq 0} Q^{(i)}$ is a disjoint union.

In this paper we shall use the following results.

Lemma 1 (Shyr, H. J. and Thierrin, G. [5]**)** *Let $i \geq 1$ and $uv \in \{p^i \mid p \in Q\}$. Then $vu \in \{p^i \mid p \in Q\}$, too. In other words, the sets $\{p^i \mid p \in Q\}$ $(i \geq 1)$ are closed under cyclic permutations of words.*

Lemma 2 (Lyndon, R. C. and Schützenberger [2]**)** *Let $f, g \in Q, f \neq g$. Then $f^m g^n \in Q$ for all $m \geq 2, n \geq 2$.*

Lemma 3 (see, e.g., **Hopcroft, J. E. and Ullman J. D** [3], p. 56, Lemma 3.1.**)** *Let L be a regular language. Then there is a constant n such that if z is any word in L and $|z| \geq n$, then we may write $z = uvw$ in such a way that $|uv| \leq n, |v| \geq 1$, and for all $i \geq 0$, $uv^i w$ is in L. Furthermore, n is not greater than the number of states of the smallest finite automaton accepting L.*

Lemma 4 (see, e.g., **Hopcroft, J. E. and Ullman J. D** [3], p. 143, Exercise 6.11.**)** *If L is a linear language then there is a constant n such that if z in L is of length n or greater, then we may write $z = uvwxy$, so that $|uvxy| \leq n, |vx| \geq 1$, and for all $i \geq 0$, $uv^i wx^i y$ is in L.*

3. Languages Consisting of Non-Primitive Words

It is easy to show that using an appropriate (deterministic) linear bounded automaton it can be decided whether a word is primitive. It is also easy to prove that for every positive integer $i > 1$ it can be decided by a (deterministic) linear bounded automaton whether a word is in $Q^{(i)}$. Similarly, by a (deterministic) linear bounded automaton it can be decided whether a word is in $\cup_{i>1} Q^{(i)}$. Thus these languages are context-sensitive. On the other side, it is well-known that the class of context-sensitive languages is closed under intersection. Thus for every context-sensitive language L, the languages $L \cap Q, L \cap Q^{(i)}, i > 1, L \cap (\cup_{i>1} Q^{(i)})$ are also context-sensitive. By this simple observation we obtain that all context-sensitive languages consisting of non-primitive words have the form $L = L' \cap (\cup_{i>1} Q^{(i)})$, where L' is context-sensitive. Conversely, for every context-sensitive language L', the language $L = L' \cap (\cup_{i>1} Q^{(i)})$ is also context-sensitive. Thus one can obtain a rather simple characterization of context-sensitive (or more complex) languages consisting of non-primitive words.
For the context-free case the following characterization is proved.

Theorem 1 (Ito, M. and Katsura, M. [1]) *Let L be a context-free language such that $L \subseteq X^+ \setminus Q$. Then $L_1 = L \cap Q^{(2)}$ is a context-free language and $L_2 = L \cap (\cup_{i\geq 3} Q^{(i)})$ is a regular language. More exactly,*

$$L_1 = F_1 \cup (\cup_{1\leq i\leq r}\{(a_i^n b_i a_i^m)^2 \mid n, m \geq 1\}) \cup (\cup_{1\leq j\leq s}\{(f_j g_j^n h_j)^2 \mid n \geq 1\})$$

where F_1 is a finite subset of $Q^{(2)}$ and $a_i^2 b_i \in Q, 1 \leq i \leq r, f_j g_j h_j \in Q, 1 \leq j \leq s$,

$$L_2 = F_2 \cup (\cup_{1\leq i\leq r} f_i^{m_i}(f_i^{k_i})^*)$$

where F_2 is a finite subset of $\cup_{i\geq 3} Q^{(i)}$ and $f_i \in Q, m_i \geq 3, k_i \geq 1, 1 \leq i \leq r$.

We now show the following

Theorem 2 *Let L be a linear language such that $L \subseteq X^+ \setminus Q$. Then $L_1 = L \cap Q^{(2)}$ is a linear language and $L_2 = L \cap (\cup_{i\geq 3} Q^{(i)})$ is a regular language. More exactly,*

$$L_1 = F_1 \cup (\cup_{1\leq j\leq s}\{(f_j g_j^n h_j)^2 \mid n \geq 1\})$$

where F_1 is a finite subset of $Q^{(2)}$, $f_j g_j h_j \in Q, 1 \leq j \leq s$, and L_2 has the same structure as in Theorem 1.

Proof: It is well-known that for every linear language L, the language $L \cap R$ is also linear whenever R is regular. On the other hand, by Theorem 1 $L_2 = L \cap (\cup_{i\geq 3} Q^{(i)})$ is regular for every context-free language L. Thus L_2 is also regular if L is linear. Therefore, $L_1 = L \cap Q^{(2)}$ is linear if L is linear. It remains to prove that $L_1 = F_1 \cup (\cup_{1\leq j\leq s}\{(f_j g_j^n h_j)^2 \mid n \geq 1\})$ where F_1 is a finite set and $f_j g_j h_j \in Q, 1 \leq j \leq s$. Observe that a language $\{(fg^n h)^2 \mid n \geq 1\}$ can be generated by the linear rules $S \to fTh, T \to gTg, T \to hf$. Thus L_1 can be generated by a linear grammar. For any $a, b \in X^+$, $\{a^j b a^{j+k} b a^k \mid j < n, k \geq n\}, \{a^j b a^{j+k} b a^k \mid j \geq n, k < n\}$ have the form $\{(fg^n h)^2 \mid n \geq 1\}$ by $f = a^j b, g = a, h = \lambda$ and $f = \lambda, g = a, h = ba^k$, in order. Moreover, $\{a^j b a^{j+k} b a^k \mid j, k < n\}$ is finite. The complement of the union of these languages is $\{a^j b a^{j+k} b a^k \mid j, k \geq n\}$. By Theorem 1, it is now enough to show that for every linear language L there exists an $n \geq 1$ such that L does not contain elements of $\{a^j b a^{j+k} b a^k \mid j, k \geq n\}$.

Assume the contrary, and let L be a linear language with a positive integer n as in Lemma 4. Using again that the intersection of a linear and a regular language is also linear, we may consider $L \cap a^+ b a^+ b a^+$ as a linear language. Then $L \cap \{a^j b a^{j+k} b a^k \mid j, k \geq 1\} \subseteq Q^{(2)}$ is also linear.

Consider an element $z = a^s b a^{s+t} b a^t$ of the language $L \cap \{a^j b a^{j+k} b a^k \mid j, k \geq 1\}$ such that $s, t \geq n$. Then, by Lemma 4, there exists a decomposition $z = uvwxy, |uvxy| \leq n, |vx| \geq 1$ such that for all $i \geq 0$, $uv^i wx^i y$ is in L. On the other hand, linear languages are closed under homomorphism and inverse homomorphism. Thus we can suppose that $u, v, x, y \in a^*$. Indeed, we can take a homomorphism $\psi : \{c, d\} \to X^*$ with $\psi(c) = a, \psi(d) = b$ and a decomposition $u'v'w'x'y'$ of the word $z' = c^s d c^{s+t} d c^t$ with $|u'v'x'y'| \leq n', |v'x'| \geq 1$, $u'v'^i w'x'^i y' \in \{c^j d c^{j+k} d c^k \mid j, k \geq 1\}, i \geq 0$, where n' is an appropriate positive integer. Obviously, we now have $\psi(u') = u, \psi(v') = v, \psi(w') = w, \psi(x') = x, \psi(y') = y$ leading to $u, v, x, y \in a^*$ if s and t are big enough.

At the same time, $z = a^s b a^{s+t} b a^t \in Q^{(2)}$ is supposed. Therefore, $a^s b \in Q$. Put $f = a^s b$ and $g = h$ with $h \in Q, h^k = a$. By Lemma 2, $f^m g^n \in Q, m, n \geq 2$.

Thus we obtain $wx^iyuv^i \in Q, i > 1$. But, by Lemma 1, $uv^iwx^iy \in Q, i > 1$ contradicting $uv^iwx^iy \in Q^{(2)}, i \geq 0$. □

Next we prove

Theorem 3 *Let L be a regular language such that $L \subseteq X^+ \setminus Q$. Then $L_1 = L \cap Q^{(2)}$ is a finite language and $L_2 = L \cap (\cup_{i \geq 3} Q^{(i)})$ has the same structure as in Theorem 1.*

Proof: By Theorem 2 it is enough to prove that for any $a, b, c \in X^+$, $L \cap Q^{(2)}$ does not contain infinitely-many elements of $\{(ab^mc)^2 \mid m \geq 1\}$. Suppose the contrary. Then, by Lemma 3, there exists a positive integer n such that $m > n$ implies a decomposition $z = uvw$ of $(ab^mc)^2$ such that $|uv| \leq n, |v| \geq 1$ and $uv^iw \in L, i \geq 0$. Regular languages are also closed under homomorphism and inverse homomorphism. Hence, similarly to the proof of the previous theorem, we may assume:

1. $v = a$ or
2. $v \in b^+$ or
3. $v \in ab^+$.

1. First suppose $v = a$. Then $a^ib^mcab^mc \in L, i \geq 0$ by Lemma 3. Using Lemma 2, by $f = ab^mc$ and $g = d, d \in Q, d^k = a$ it holds that $ab^mcab^mca^i \in Q, i \geq 2$. Applying Lemma 1 we obtain $a^ib^mcab^mc \in Q, i \geq 3$ contradicting $a^ib^mcab^mc \in Q^{(2)}, i \geq 0$.

2. Suppose $v \in b^+$. Then there exists $k > 0$ such that $ab^{m+i*k}cab^mc \in L, i \geq 0$ by Lemma 3. Using Lemma 2, by $f = cab^m$ and $g = d, d \in Q, d^j = b$ it holds that $cab^mcab^{m+i} \in Q, i \geq 2$. Applying Lemma 1 we obtain $ab^{m+i}cab^mc \in Q, i \geq 2$ contradicting $ab^{m+i*k}cab^mc \in Q^{(2)}, i \geq 0$.

3. Finally suppose $v \in ab^+$. Then there exists $k > 0$ such that $(ab^k)^ib^{m-k}cab^mc \in L, i \geq 0$ by Lemma 3. Using Lemma 2, by $f = b^{m-k}cab^k$ and $g = d, d \in Q, d^j = ab^k$ it holds that $b^{m-k}cab^mcab^k(ab^k)^i \in Q, i \geq 2$. Applying Lemma 1 we obtain $(ab^k)^{i+1}b^{m-k}cab^mc \in Q, i \geq 2$ contradicting $(ab^k)^ib^{m-k}cab^mc \in Q^{(2)}, i \geq 0$. □

4. Summary

We can summarize our results in the following.

Corollary 1 *Let $\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$ be classes of languages such that $\mathcal{L}_0$ is the class of all finite languages in $X^+ \setminus Q$, $\mathcal{L}_1$ is the class of languages having the form $\cup_{1 \leq i \leq r} f_i^{m_i}(f_i^{k_i})^*$, $f_i \in Q$, $m_i \geq 3, k_i \geq 1, 1 \leq i \leq r$, $\mathcal{L}_2$ is the class of languages of the form $\cup_{1 \leq j \leq s}\{(f_jg_j^nh_j)^2 \mid n \geq 1\}$, $f_jg_jh_j \in Q, 1 \leq j \leq s$, and $\mathcal{L}_3$ is the class of languages with the structure $\cup_{1 \leq i \leq r}\{(a_i^nb_ia_i^m)^2 \mid n, m \geq 1\}$ where $a_i^2b_i \in Q, 1 \leq i \leq r$. Then the following statements hold:*

(a) L is a context-free language consisting of non-primitive words if and only if
$L = L_0 \cup L_1 \cup L_2 \cup L_3, L_i \in \mathcal{L}_i, i = 0, 1, 2, 3$.

(b) L is a linear language consisting of non-primitive words if and only if
$L = L_0 \cup L_1 \cup L_2, L_i \in \mathcal{L}_i, i = 0, 1, 2$.

(c) L is a regular language consisting of non-primitive words if and only if
$L = L_0 \cup L_1, L_i \in \mathcal{L}_i, i = 0, 1$.

(d) L is a finite language consisting of non-primitive words if and only if
$L = L_0, L_0 \in \mathcal{L}_0$.

Of course, the statement (a) is the same as Theorem 1 [1] and the statement (d) is trivial.

References

[1] Ito, M., Katsura, M., Context-free languages consisting of non-primitive words, *Semigroup Forum* 37 (1988), 45–52.

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

[3] Hopcroft, J. E., Ullman, J. D., Introduction to Automata Theory, languages, and Computation, *Addison-Wesley, Reading, Mass.,* 1979.

[4] Shyr, H. J., Free Monoids and Languages, 2nd ed., *Lecture Notes, Inst. of Applied Math., National Chung-Hsing Univ., Taichung, Taiwan, R.O.C.,* 1991.

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