공학 연구논문집 제23권 제2호 pp. 81~93, 울산대학교
Journal of Engineering Research Vol. 23, No. 1, pp. 81~93, 1992.
University of Ulsan.

LR(k) 파서의 동작에 대한 문법적 묘사에 관한 연구*

이명준 · 박양수
전자계산학과

요 약

임의의 문맥자유문법이 주어졌을 때 이 문법에 대한 LR(k) 파서의 동작을 문법기호를 이용하여 추상적으로 묘사할 수 있다는 것을 밝혔다. 이를 위하여, 임의의 입력문장을 처리하기 위한 LR(k) 파서의 동작 과정을 자신의 문법문장으로 표현할 수 있는 새로운 LR(k) 기계묘사문법을 제안하였다.

Grammatical Description of the Behavior of LR(k) Parsers*

Myung-Joon Lee · Yang-Su Park
Department of Computer Science

Abstract

We show that given a context-free grammar G, the behavior of the LR(k) parser for G can be abstractly described in terms of grammar symbols. For this, we introduce an *LR(k)* machine description grammar whose sentences describe the sequences of actions taken by a given LR(k) parser.

* Research supported in part by NON DIRECTED RESEARCH FUND, Korea Research Foundation 1989.

1. Introduction

Since the announcement of LR(k) grammars and their parsing[6], much work has been done to discover the properties of LR(k) grammars and LR(k) parsing[2,4,5,8]. Recently, a new grammar called *LR(k)-colored grammar*[7] has been introduced as a tool for transforming an LR(k) grammar into an SLR(k) grammar[3] that covers the LR(k) grammar. Besides the covering property, the LR(k)-colored grammar can be effectively used for establishing formal properties related with LR(k) parsing since the grammar transforms each reduction of an LR(k) parser into its grammar symbol in addition to the transformation of each *GOTO* transition.

In this paper, we exploit the properties of the LR(k)-colored grammar useful for describing the behavior of a given LR(k) parser by showing that there is a one-to-one correspondence between *shift* and/or *reduce* actions of the parser, and those symbols that appear in the LR(k)-colored grammar. Using the properties, we present an *LR(k) machine description grammar* whose sentences describe actions of a given LR(k) parser abstractly.

The organization of this paper is as follows. In Section 2, the notion of LR(k) parsing is revisited for clarifying the arguments in the later sections after the notation and some definitions are given. In Section 3, LR(k)-colored grammar is revisited and its fundamental properties are examined. In Section 4, an *LR(k) machine description grammar* is presented as a result of this paper after the descriptive power of LR(k)-colored grammar is exploited.

2. Notation and Definitions

This section reviews the basic concepts concerning context-free grammars and LR(k) parsing, and introduces some fundamental definitions and known results with our notation. For general background the reader is refered to [1].

A *context-free grammar* (CFG) is a 4-tuple $G=(N, \Sigma, P, S)$, where N is a finite set of nonterminals; Σ is a finite set of *terminals* such that $N \cap \Sigma = \varnothing$; P is a finite subset of $N \times V^*$, where V(vocabulary) stand for $N \cup \Sigma$, and each member(A,α) of P is called a *production*, written $A \rightarrow \alpha$, and the Greek letter π is reserved to denote a production; and S is the *start symbol*. For the convenient description of LR(k) parsing, G is assumed to be *augmented* in the sense that P contains a special (start) production $S' \rightarrow S$, where S' does not occur in any other procuction.

Lower-case Greek letters, such as α,β, and γ denote vocabulary stings in V^*; lower-case Roman letters near the beginning of the alphabet, such as a, b, and c are terminals in Σ, and those near the end, such as w,x,y, and z are terminal strings in Σ^*; upper-case Roman letters near the end of the alphabet, such as W,X, and Y are vocabulary symbols in V. The *empty string* is denoted as ε, and is of length 0. For two vocabularies V_1 and V_2, a *homomorphism* is defined as a mapping h: $V_1 \rightarrow V_2^*$; the domain of the homomorphism h, throughout the paper, is extended to V_1^* by letting $h(\varepsilon)=\varepsilon$ and $h(xX)=h(x)\cdot h(X)$ for all $x\in V_1^*$, and $X\in V_1$. The homomorphism h is said to be *fine* if $h:V_1\rightarrow V_2\cup\{\varepsilon\}$, and *very fine* if $h: V_1\rightarrow V_2$.

When a production $A\rightarrow\alpha$ is in P, a derivation $\beta A\gamma\Rightarrow\beta\alpha\gamma$ holds, and if $\gamma\in\Sigma^*$, the derivation is said to be *rightmost*, and is written $\beta A\gamma\Rightarrow_{rm}\beta\alpha\gamma$. The reflexive transitive closure, and the transitive closure of the relation $\Rightarrow$ are denoted by $\Rightarrow^*$ and $\Rightarrow^+$, respectively. A *sentential form* of G is a string α such that $S\Rightarrow^*\alpha$, and a *right sentential form* is a string α such that $S\Rightarrow^*_{rm}\alpha$. The *language generated by* a string α is $L(\alpha)=\{w\in\Sigma^*\mid\alpha\Rightarrow^*w\}$. Conventionally, $L(G)$ is used to denote $L(S)$. A string γ is called a *viable prefix* of G if γ is a prefix of $\alpha\beta$ such that $S'\Rightarrow^*_{rm}\alpha Aw\Rightarrow_{rm}\alpha\beta w$.

A grammar G is *ambiguous* if there exists a string in $L(G)$ with more than one rightmost derivations. G is said to be *unambiguous* if it is not ambiguous. G is *ambiguous of degree n* if every string in $L(G)$ has at most n distinct rightmost derivations. Clearly, G is ambiguous of degree 1 if and only if G is unambiguous. G is *boundedly ambiguous* if there exists some integer n such that G is ambiguous of degree n. G is said to be *unboundedly ambiguous* if it is not boundedly ambiguous.

A pair $[A\rightarrow\alpha.\beta,u]$ is an *LR(k) item* for G if $A\rightarrow\alpha\beta\in P$ and $u\in FIRST_k(\Sigma^*\$)$, where the special symbol \$ not in Σ denotes the *endmarker* of an input string and the function $FIRST_k$ is defined as $FIRST_k(\alpha)=\{PREF_k(x)\mid x\in L(\alpha)\}$ ($PREF_k(x)$ denotes the prefix of x of length k, or just x if the length of x is less than k). It is said to be *valid* for $\gamma\alpha$, a viable prefix of G, if G permits a derivation $S'\Rightarrow^*_{rm}\gamma Aw\Rightarrow_{rm}\gamma\alpha\beta w$, for which $k=PREF_k(w\$)$. In an item, the left part of the comma is called the *core* of the item, and the right part is called the *lookahead* of the item. The function *closure*, which maps a set of LR(k) items to another set of LR(k) items, is defined recursively as follows. Let q be a set of LR(k) items. Then

$$closure(q)=_s q\cup\{[B\rightarrow.\gamma,v]\mid[A\rightarrow\alpha.B\beta,u]\in closure(q),\ v\in FIRST_k(\beta u),\ B\rightarrow\gamma\in P\},$$

where the notation "$X=_s f(X)$" means that X is the smallest set which satisfies the condition $X=f(X)$. The canonical collection of sets of LR(k) items for G, denoted C_k, is defined recursively by

$$C_k=_s\{q_0\}\cup\{GOTO(q,X)\mid q\in C_k, X\in V\},$$

where $q_0=closure(\{[S'\rightarrow.S,\$]\})$, and

$$GOTO(q,X)=closure(\{[A\rightarrow\alpha X.\beta,u]\mid[A\rightarrow\alpha.X\beta,u]\in q\}).$$

An element of C_k is said to be an LR(k) *state* over G. We call the state q_0 the *initial state*. The domain of the *GOTO* function is extended to $C_k\times V^*$ as follows:

$$GOTO(q,\varepsilon)=q \text{ and } GOTO(q,X\gamma)=GOTO(GOTO(q,X),\gamma).$$

To clarify the arguments in this paper, a nondeterministic LR(k) parser for the whole class of context-free grammars is defined by the following formal system called *LR(k) machine*. The *LR(k) machine* for G is a 4-tuple $LRM_k(G)=(C_k, GOTO\ ACTION, q_0)$, where C_k, *GOTO*, and q_0 are as was stated above; *ACTION* is a function from $C_k\times$

$FIRST_k(\Sigma^* \$)$ to subsets of $\{shift, accept\} \cup \{\text{reduce } \pi \mid \pi \in P\}$ defined by

$ACTION(q,u) = \{shift\}$ if $[A \to \alpha.\beta,v] \in q$, $A \neq S'$, $\beta \neq \varepsilon$, and $u \in EFF_k(\beta v)$
$\cup \{reduce\ \pi\}$ if π is $A \to \alpha$, $[A \to \alpha.,u] \in q$, and $A \neq S'$
$\cup \{\text{accept}\}$ if $[S' \to S., \$] \in q$, and $u = \$$,

where EFF_k, the ε-free $FIRST_k$ function, is

$EFF_k(\alpha) = \{PREF_k(\alpha) \mid \alpha \in \Sigma^* \$\} \cup \{PREF_k(x) \mid \alpha \Rightarrow^*_{rm} \beta \Rightarrow_{rm} x,\ \beta \neq Ax$ for *all* $A \in N$, $x \in \Sigma^* \$\}$.

It would be noted that if $ACTION(q,u)$ contains *shift*, then the state q contains an LR(k) item $[A \to \alpha.a\beta,v]$ with $u \in FIRST_k(a\beta v)$ because of the definition of the *closure* function and the EFF_k function above.

A *configuration* of $LRM_k(G)$ is a triple $(\sigma, w\$, \Pi)$ in $C_k^* \times \Sigma^*\$ \times P^*$, where σ represents the *state stack* as a sequence of LR(k) states, w the *remaining input string*, and Π the *current output right parse*. The *initial configuration* for an input string z in Σ^* is $(q_0, z\$, \varepsilon)$. We define the *state stack induced by a viable prefix* γ, denoted σ_γ, by

$\sigma_\gamma = q_0$ if $\gamma = \varepsilon$; otherwise $\sigma_\gamma = \sigma_{\beta X} = \sigma_\beta\ GOTO(q_0, \beta X)$ with $\gamma = \beta X$

(obviously, $\sigma_\gamma = \sigma_\beta$ if and only if $\gamma = \beta$ because the *GOTO* function is deterministic). The *top* of the state stack σ_γ, denoted $top(\sigma_\gamma)$, is the state $GOTO(q_0,\gamma)$. A move by $LRM_k(G)$ is represented by the following binary relations on configurations.

(1) $\vdash_{shift}$ relation (*shift* move), defined by $(\sigma_\gamma, aw\$, \Pi) \vdash_{shift} (\sigma_{\gamma a}, w\$, \Pi)$ if $ACTION(top(\sigma_\gamma)), PREF_k(aw\$)$ contains *shift*;

(2) $\vdash_\pi$ relation for $\pi \in P$ (*reduce* π move), defined by $(\sigma_{\gamma\alpha}, w\$, \Pi) \vdash_\pi (\sigma_{\gamma A}, w\$, \Pi\pi)$ if $ACTION(top(\sigma_{\gamma\alpha}), PREF_k(w\$)))$ contains *reduce* π with π being $A \to \alpha \in P$;

(3) $\vdash_{accept}$ relation (*accept* move), defined by $(\sigma_\gamma, w\$, \Pi) \vdash_{accept} (\varepsilon, \$, \Pi)$ if $ACTION(top(\sigma_\gamma), PREF_k(w\$))$ contains *accept*

(In this case, $w = \varepsilon$, and $\sigma_\gamma = \sigma_S = q_0 \cdot GOTO(q_0,S)$).

For configurations C_1 and C_2, we write $C_1 \vdash C_2$ if at least one of the relations defined above holds between C_1 and C_2. In particular, we write $(\sigma_\gamma, w\$, \Pi) \vdash$ *error* if $ACTION(top(\sigma_\gamma), PREF_k(w\$))$ is $\varnothing$. The *language accepted by* $LRM_k(G)$, denoted $L(LRM_k(G))$, which is equivalent to $L(G)$, is defined by

$L(LRM_k(G)) = \{z \in \Sigma^* \mid (q_0, z\$, \varepsilon) \vdash^+ (\varepsilon, \$, \Pi)$ for some $\Pi \in P^+\}$

In this, Π is *right parse* of z. We use the following terminology for configurations of $LRM_k(G)$. Let C be a configuration. (1) C is *valid for a string* z in Σ^* if $(q_0, z\$, \varepsilon) \vdash^* C$. If C is valid for some string in Σ^*, then C is a *valid configuration of* $LRM_k(G)$. (2) C is *acceptable* if C is valid and $C \vdash^* (\varepsilon, \$, \Pi)$ for some Π in P^+. (3) C is *nondeterministic* if $C \vdash C_1$, $C \vdash C_2$, and $C_1 \neq C_2$. Here, each of the moves from C is also said to be *nondetermin-istic*.

$LRM_k(G)$ is said to be deterministic if no (valid) configurations of $LRM_k(G)$ are nondeterministic. Obviously, $LRM_k(G)$ is deterministic iff $ACTION_k(q,u)$ has at most one element for any $q \in C_k$, and $u \in FIRST_k($

$\Sigma^*\$$)(i.e., G is LR(k)). Let $(\sigma_\gamma, w\$, \Pi)$ be a valid configuration for a string z in Σ^*. We say that a sequence of moves from the initial configuration $(q_0, z\$, \varepsilon)$ to the valid configuration is an *LR(k) parsing subsequence* of z over G. If $\sigma_\gamma = \varepsilon$, it is said to be a *valid LR(k) parsing sequence* of z, whereas it is said to be *invalid* if $(\sigma_\gamma, w\$, \Pi) \vdash$ *error*. It would be worth pointing out that if $LRM_k(G)$ is not deterministic, then there exists a string in $L(G)$ with multiple distinct LR(k) parsing sequences whether they are valid or not.

This section is ended by recalling some well-known properties of LR(k) parsing which are fundamental to the arguments in the remaining sections.

Property 2.1. *$GOTO(q_0,\gamma)$ is defined if and only if γ is a viable prefix of G.*

Property 2.2. *C_k is equivalent* to *Φ_k, the collection of the sets of valid LR(k) items for G, defined by*

$\Phi_k = \{q \mid q$ is the set of valid LR(k) items for some viable prefix of $G\}$.

Moreover, a state q is the set of valid LR(k) items for a viable prefix γ iff $q = GOTO(q_0, \gamma)$.

Property 2.3. *There is a derivation in G such that*

$S' \overset{*}{\Rightarrow}_{rm} \gamma_1 w_1 \overset{*}{\Rightarrow}_{rm} \gamma_2 w_2 w_1 \Rightarrow^* z$, and γ_1 and γ_2 *are viable prefixes of G*

if and only if there is a valid LR(k) parsing sequence over G such that

$(q_0, z\$, \varepsilon) \vdash^* (\sigma_{\gamma 2}, w_2 w_1\$, \Pi) \vdash^* (\sigma_{\gamma 1}, w_1\$, \Pi') \vdash^* (\varepsilon, \$, \Pi'')$ for *some* $\Pi, \Pi', \Pi'' \in P^*$.

Property 2.4. *If there is an invalid LR(k) parsing sequence of xw over G such that*

$(q_0, xw\$, \varepsilon) \vdash^* (\sigma_\gamma, w\$, \Pi) \vdash$ *error,*

then there is a string y in Σ^* such that

$(q_0, xy\$, \varepsilon) \vdash^* (\sigma_\gamma, y\$, \Pi) \vdash^* (\varepsilon, \$, \Pi')$.

That is, every invalid LR(k) parsing sequence of a string is a subsequence of a valid LR(k) parsing sequence of another string in L(G).

3. LR(k)-Colored Grammar

In this section, we revisit *LR(k)-colored grammar*[7] which is constructed from LR(k) machine on the ground that *GOTO* transitions on terminal symbols from each LR(k) state causes *shift* moves, and reducible items in the state causes *reduce* moves on the productions in the items, while *GOTO* transitions on nonterminal symbols causes state transitions as a part of *reduce* moves. For representing such moves of the LR(k) machine as grammar symbols, we transfigure each *GOTO* transition into a symbol of the introduced grammar, and also transfigure each reduction. As a result, we will have a one-to-one correspondence between *shift* and/or *reduce* moves of the LR(k) machine, and those symbols that appear in the LR(k)-colored grammar;

Defining the LR(k)-colored grammar, we use new notations X^q and π^q to denote nonterminals of the grammar; they mean the transition on symbol X from state q and the reduction on π at state q, respectively.

Construction 3.1. Let a CFG $G = (N, \Sigma, P, S)$, and $LRM_k(G) = (C_k, GOTO, ACTION, q_0)$

The *LR(k)-colored grammar* for G is $\mathbf{G} = (\mathbf{N}, \mathbf{\Sigma}, \mathbf{P}, \mathbf{S})$, where

(1) $\mathbf{N} = \{S\} \cup \{X^q \mid q \in C_k, [A \to \alpha . X\beta, u] \in q, X \in V\}$

$\cup \{\pi^q \mid q \in C_k,\ [A \to \alpha.,u] \in q,\ A \neq S',\ \pi \text{ is } A \to \alpha \in P\}$,

The set of new vocabularies, $N \cup \Sigma$, is denoted by V. For notational convenience, we classify N into four disjoint sets, $\{S\}$, N_N, N_Σ, and N_P, as follows:

$N_N = \{A^q \mid A \in N,\ A^q \in N\}$, $N_\Sigma = \{a^q \mid a \in \Sigma,\ a^q \in N\}$, $N_P = \{\pi^q \mid \pi \in P,\ \pi^q \in N\}$; and the set $N_N \cup N_\Sigma$ is denoted by N_V.

(2) $P = \{S \to S^{q_0}\}$
$\cup \bigcup_{q \in C_k} \{A^q \to \theta(q,\alpha) \cdot \pi^{q'} \mid A \neq S'\ [A \to .\alpha, u] \in q, \pi \text{ is } A \to \alpha,\ q' = GOTO(q,\alpha)\}$
$\cup \bigcup_{a^q \in N_\Sigma} \{a^q \to a\} \cup \bigcup_{\pi^q \in N_P} \{\pi^q \to \varepsilon\}$,

where $\theta(q,\alpha)$ is a function from $C_k \times V^*$ to N_V^* defined by

$\theta(q,\varepsilon) = \varepsilon$ and $\theta(q, X\alpha) = X^q \cdot \theta(GOTO(q,X),\alpha)$

if X^q is in N_V (or equivalently $GOTO(q,X)$ is in C_k).

Example 3.1. Consider an unambiguous context-free grammar $G = (\{S, A, B\}, \{a, b, 0, 1\}, P, S)$ with P:

$\pi_1: S \to Aa$, $\pi_2: S \to Bb$, $\pi_3: A \to 0A\ 1$, $\pi_4: A \to 0\ 1$, $\pi_5: B \to 0B\ 11$, $\pi_6: B \to 011$

By Construction 3.1 and the LR(1) machine for G' exposed in Figure 3.1, we get the LR(1)-colored grammar $\boldsymbol{G} = (\boldsymbol{N}, \Sigma, \boldsymbol{P}, S)$, where

(1) $\boldsymbol{N} = \{S\} \cup N_N \cup N_\Sigma \cup N_P$, where

$N_N = \{S^{q_0}, A^{q_0}, A^{q_6}, A^{q_{14}}, B^{q_0}, B^{q_6}, B^{q_{14}}\}$

$N_\Sigma = \{a^{q_2}, b^{q_4}, 0^{q_0}, 0^{q_6}, 0^{q_{14}}, 1^{q_6}, 1^{q_7}, 1^{q_9}, 1^{q_{10}}, 1^{q_{12}}, 1^{q_{14}}, 1^{q_{15}}, 1^{q_{17}}, 1^{q_{18}}\ 1^{q_{20}}\}$

$N_P = \{\pi_1^{q_3}\ \ \pi_2^{q_5}\ \ \pi_3^{q_8}\ \ \pi_3^{q_{16}}\ \ \pi_4^{q_{12}}\ \ \pi_4^{q_{20}}\ \ \pi_5^{q_{11}}\ \ \pi_5^{q_{19}}\ \ \pi_6^{q_{13}}\ \ \pi_6^{q_{21}}\}$

(2) $\boldsymbol{P}$ is composed of the following productions:

$S \to S^{q_0}$,

$S^{q_0} \to A^{q_0}\ a^{q_2}\ \pi_1^{q_3} \mid B^{q_0}\ b^{q_4}\ \pi_2^{q_5}$,

$A^{0} \to 0^{q_0}\ A^{q_6}\ 1^{q_7}\ \pi_3^{q_8} \mid 0^{q_0}\ 1^{q_6}\ \pi_4^{q_{12}}$,

$B^{q_0} \to 0^{q_0}\ B^{q_6}\ 1^{q_9}\ 1^{q_{10}}\ \pi_5^{q_{11}} \mid 0^{q_0}\ 1^{q_6}\ 1^{q_{12}}\ \pi_6^{q_{13}}$,

$A^{q_6} \to 0^{q_6}\ A^{q_{14}}\ 1^{q_{15}}\ \pi_3^{q_{16}} \mid 0^{q_6}\ 1^{q_{14}}\ \pi_4^{q_{20}}$,

$B^{q_6} \to 0^{q_6}\ B^{q_{14}}\ 1^{q_{17}}\ 1^{q_{18}}\ \pi_5^{q_{19}} \mid 0^{q_6}\ 1^{q_{14}}\ 1^{q_{20}}\ \pi_6^{q_{20}}$,

$A^{q_{14}} \to 0^{q_{14}}\ A^{q_{14}}\ 1^{q_{15}}\ \pi_3^{q_{16}} \mid 0^{q_{14}}\ 1^{q_{14}}\ \pi_4^{q_{20}}$,

$B^{q_{14}} \to 0^{q_{14}}\ B^{q_{14}}\ 1^{q_{17}}\ 1^{q_{18}}\ \pi_5^{q_{19}} \mid 0^{q_{14}}\ 1^{q_{14}}\ 1^{q_{20}}\ \pi_6^{q_{21}}$,

$X^q \to X$ for all $X^q \in N_\Sigma$

$\pi^q \to \varepsilon$ for all $\pi^q \in N_P$

* The grammar G is not LR(k) for any k and the language $\{0^n 1^n a \mid n \geq 1\} \cup \{0^n 1^{2n} b \mid n \geq 1\}$ generated by G is not deterministic context-free [1].

For observing the basic properties of the LR (k)-colored grammar, we define a homomorphism h from the vocabulary symbols of **G** to those of G.

Definition 3.1. A fine homomorphism h: $V \rightarrow \mathrm{V} \cup \{\varepsilon\}$ is

$$h(X) = \begin{cases} S', \text{ if } \boldsymbol{X} = \boldsymbol{S} \\ \varepsilon, \text{ if } \boldsymbol{X} \in \boldsymbol{N}_p \\ X, \text{ if } \boldsymbol{X} = X^q \in \boldsymbol{N}_V \\ a, \text{ if } \boldsymbol{X} = \mathrm{a} \in \Sigma \end{cases}$$

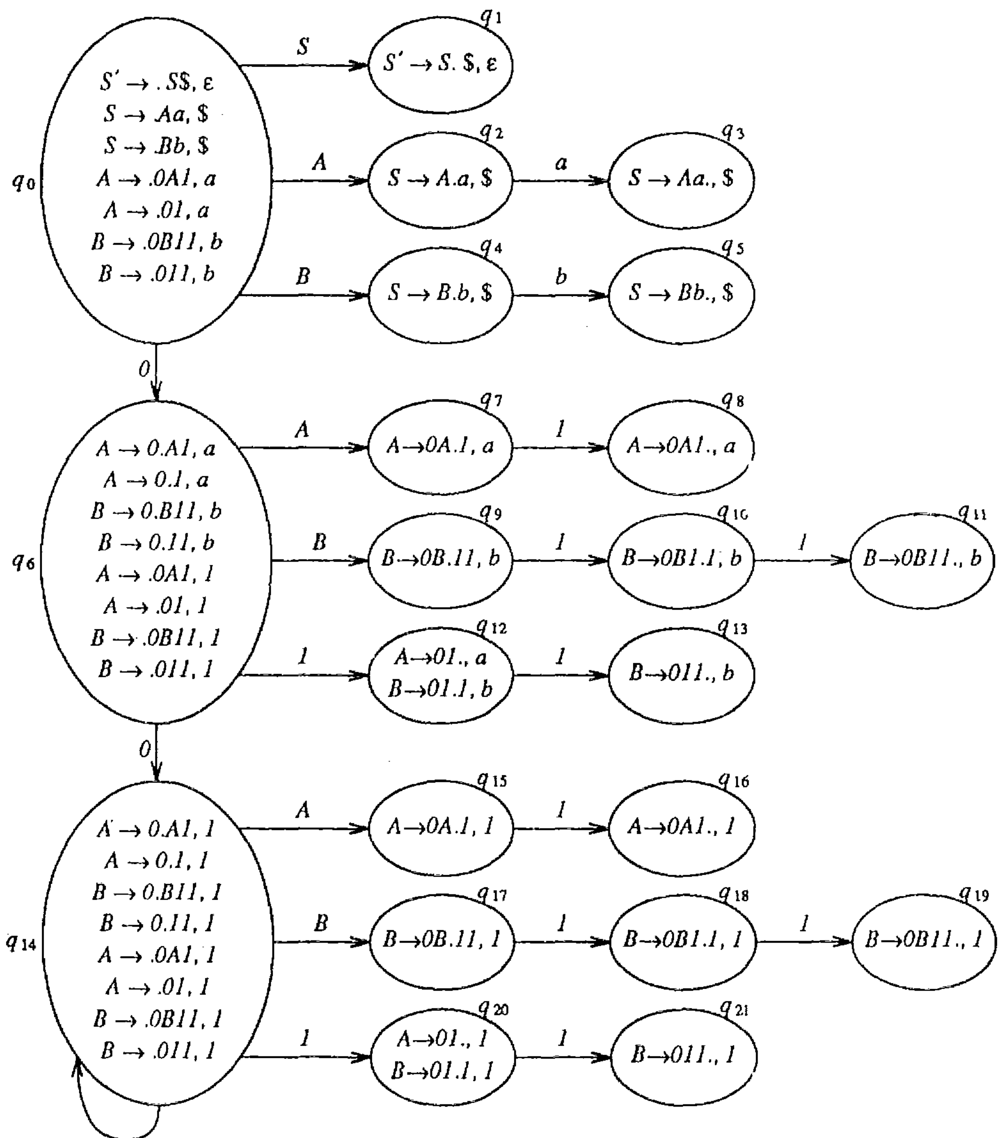

Figure 3.1. LR(1) machine for G in Example 3.1

The following properties associated with G are fundamental to our later arguments, and can be easily obtained from the definitions of the grammar G and the homomorphism h.

Property 3.1.

(1) For an arbitrary LR(k) state q, $\theta(q,\alpha)\theta(GOTO(q,\alpha),\beta)=\theta(q,\alpha\beta)$.

(2) $\theta(q_0,\gamma)$ is defined if and only if γ is a viable prefix of G.

(3) Let $\gamma_1=\theta(q_0,\gamma_1)$ and $\gamma_2=\theta(q_0,\gamma_2)$. Then there is a derivation in G such that $S \Rightarrow^*_{rm} \gamma_1 w_1 \Rightarrow^*_{rm} \gamma_2 w_2 w_1 \Rightarrow^* z$ if and only if there is a derivation in G such that $S' \Rightarrow^*_{rm} \gamma_1 w_1 \Rightarrow^*_{rm} \gamma_2 w_2 w_1 \Rightarrow^* z$.

(4) If there is a derivation in G such that $S \Rightarrow^*_{rm} \gamma X^q \alpha$, then $q=GOTO(q_0, h(\gamma))$.

A useful one-to-one correspondence between rightmost derivations in G and valid LR(k) parsing sequences over G can be established by the following theorem.

Theorem 3.1. *Let* $\gamma_1=\theta(q_0,\gamma_1)$ and $\gamma_2=\theta(q_0,\gamma_2)$. *Then there is a derivation in* G *such that*

$$S \Rightarrow^*_{rm} \gamma_1 w_1 \Rightarrow^*_{rm} \gamma_2 w_2 w_1 \Rightarrow^* z$$

if and only if there is an LR(k) parsing sequence over G such that

$$(q_0,z\$,\varepsilon) \vdash^* (\sigma_{\gamma 1} w_2 w_1\$,\Pi) \vdash^* (\sigma_{\gamma 1}, w_1\$,\Pi') \vdash^* (\varepsilon,\$,\Pi'') \text{ for } some\ \Pi,\Pi',\Pi'' \in P^*.$$

Proof. By Property 3.1-(2), γ_1 and γ_2 are viable prefixed of G. Then the theorem follows from Property 3.1-(3) and Property 2.3. □

4. Describing Moves of an LR(k) Machine

In the formal section, it was claimed that there is a one-to-one correspondence between *shift* and/or *reduce* moves of an LR(k) machine, and some kinds of symbols that appear in the related LR(k)-colored grammar G. The correspondence can be stated by the following two theorems (Theorem 4.1 and Theorem 4.2): one says that if a symbol in N_Σ appears in a sentential form of G, then the LR(k) machine can take a *shift* move corresponding to the symbol, and vice versa; the other says that if a symbol π^q in N_P appears in a sentential form of G, then the LR(k) machine can take a *reduce* move corresponding to the symbol, and vice versa. In this section, we establish the validity of the theorems, and present an LR(k) *machine description grammar* of which sentences describe parsing sequences of a given LR(k) machine.

Theorem 4.1. *Let* a^q *be a nonterminal in* N_Σ. *Then there is a derivation in* **G** *such that*

$$S \Rightarrow^*_{rm} \gamma a^q w \Rightarrow^* z$$

if and only if there is a valid LR(k) parsing sequence such that

$$(q_0,z\$,\varepsilon) \vdash^* (\sigma_\gamma,aw\$,\Pi) \vdash_{shift} (\sigma_{\gamma a},w\$,\Pi) \vdash^* (\varepsilon,\$,\Pi'),$$

where $\gamma=\theta(q_0,\gamma)$, *top* $\sigma_\gamma=q$, *and* $\Pi,\Pi', \in P^*$.

Theorem 4.2. *Let* π^q *be a nonterminal in* N_P *with* π *being* $A\rightarrow\gamma_2$. *Then, there is a derivation in* **G** *such that*

$$S \Rightarrow^*_{rm} \gamma \pi^q w \Rightarrow^* z$$

if and only if there is a valid LR(k) parsing sequence such that

$(q_0,z\$,\varepsilon) \vdash^* (\sigma_{\gamma_1\gamma_2},w\$,\Pi) \vdash_\pi (\sigma_{\gamma_1 A},w\$,\Pi\pi) \vdash^* (\varepsilon,\$,\Pi')$,

where $\gamma=\theta(q_0,\sigma\gamma_1\gamma_2)$, *top* $(\sigma_{\gamma_1\gamma_2})=q$, *and* $\Pi, \Pi' \in P^*$.

Theorem 4.1 says that if a symbol in N_Σ appears in a sentential form of G, then the LR(k) machine can take a *shift* move corresponding to the symbol, and vice versa.

Theorem 4.2 says that if a symbol π^q in N_P appears in a sentential form of G then the LR (k) machine can take a *reduce* move corresponding to the symbol, and vice versa. This one-to-one correspondence can be appreciated by the following example.

Example 4.1. For the grammar G in Example 3.1 and a sentence of G, z=0011a, consider the followings.

LR(k) Parsing Sequence of z over G

$$
\begin{array}{lll}
(q_0,0011a\$,\varepsilon) & \vdash_{shift} & (q_0q_6,011a\$,\varepsilon) \\
 & \vdash_{shift} & (q_0q_6q_{14},11a\$,\varepsilon) \\
 & \vdash_{shift} & (q_0q_6q_{14}q_{20},1a\$,\varepsilon) \\
 & \vdash_{\pi_4} & (q_0q_6q_7,1a\$,\pi_4) \\
 & \vdash_{shift} & (q_0q_6q_7q_8,a\$,\pi_4) \\
 & \vdash_{\pi_3} & (q_0q_2,a\$,\pi_4\pi_3) \\
 & \vdash_{shift} & (q_0q_2q_3,\$,\pi_4\pi_3) \\
 & \vdash_{\pi_1} & (q_0q_1,\$,\pi_4\pi_3\pi_1) \\
 & \vdash_{accept} & (\varepsilon,\$,\pi_4\pi_3\pi_1)
\end{array}
$$

Rightmost Derivation for z in G

$$
\begin{array}{lll}
0011a & \Leftarrow_{rm} & 0^{q_0}011a \\
 & \Leftarrow_{rm} & 0^{q_0}0^{q_6}11a \\
 & \Leftarrow_{rm} & 0^{q_0}0^{q_6}1^{q_{14}}1a \\
 & \Leftarrow_{rm} & 0^{q_0}0^{q_6}1^{q_{14}}\,\pi_4^{q_{20}}1a \Leftarrow_{rm} 0^{q_0}A^{q_6}1a \\
 & \Leftarrow_{rm} & 0^{q_0}A^{q_6}1^{q_7}a \\
 & \Leftarrow_{rm} & 0^{q_0}A^{q_6}1^{q_7}\,\pi_3^{q_8}\,a \Leftarrow_{rm} A^{q_0}a \\
 & \Leftarrow_{rm} & A^{q_0}a^{q_2} \\
 & \Leftarrow_{rm} & A^{q_0}a^{q_2}\,\pi_1^{q_3} \quad \Leftarrow_{rm} S^{q_0} \\
 & \Leftarrow_{rm} & S
\end{array}
$$

The reversed sequence of the rightmost nonterminals appearing in the above rightmost derivation, contained in $N_\Sigma \cup N_P$, is $0^{q_0}0^{q_6}1^{q_{14}}$ $\pi_4^{q_{20}}$ 1^{q_7} $\pi_3^{q_8}$ a^{q_2} $\pi_1^{q_3}$. Note that the symbols in this sequence are in one-to-one correspondence with the moves of the LR(k) parsing sequence.

For clarifying the property formally, we remark the following. Let X_i' s be symbols in V, and α_j' s be strings in V^* such that $X_i \Rightarrow^* x_i$ for $1 \le i \le n$, and $\alpha_j \Rightarrow^* w_j$ for $1 \le j \le n+1$. Then, the following three statements are equivalent.

(1) G permits a derivation

$$S \Rightarrow^* \alpha_1X_1\alpha_2X_2\cdots\alpha_nX_n\alpha_{n+1} \Rightarrow^* z$$
$$(=w_1x_1w_2\,x_2\cdots w_{n+1})$$

(2) G permits a rightmost derivation

$$S \Rightarrow^*_{rm}\gamma_nX_nw_{n+1} \Rightarrow^*_{rm}\cdots\Rightarrow^*_{rm}\gamma_2X_2w_3x_3\cdots x_nw_{n+1} \Rightarrow^*_{rm}\gamma_1X_1w_2\cdots x_nw_{n+1} \Rightarrow^* z,$$

where $\gamma_i \Rightarrow^* \alpha_1X_1\cdots\alpha_{i-1}X_{i-1}\alpha_i$ for $1 \le i \le n$.

(3) there is a valid LR(k) parsing sequence over G such that

$$
\begin{array}{ll}
(q_0,z\$,\varepsilon) & \vdash^* (\sigma_{\gamma_1},x_1w_2\cdots x_nw_{n+1}\$,\Pi_1) \\
 & \vdash^* (\sigma_{\gamma_2},x_2w_3\cdots x_nw_{n+1}\$,\Pi_2) \\
 & \cdots \\
 & \vdash^* (\sigma_{\gamma_n},x_nw_{n+1}\$,\Pi_n) \\
 & \vdash^* (\varepsilon,\$,\Pi)
\end{array}
$$

where $\gamma_i=\theta(q_0, \gamma_i)$ and where $\gamma_i \Rightarrow^* \alpha_1X_1\cdots\alpha_{i-1}X_{i-1}\alpha_i$ for $1 \le i \le n$ (by Theorem 3.1).

For the formal proofs of Theorem 4.1 and 4.2, we will examine the relationship between right sentential forms of G and acceptable configurations of the LR(k) machine, and the relationship between right sentential forms of G and moves of the machine in order. First, we introduce some relations which are useful for capturing the relationship.

Definition 4.1. Two relations *sdescribes* (stands for "describes a *shift* move") and *rdescribes* (stands for "describes a *reduce* move") from right sentential forms of G to acceptable configurations of $LRM_k(G)$ are defined as follows:

(1) Let a^q be a nonterminal in N_Σ, and $\gamma a^q w$ a right sentential form of G. Then,

$\gamma a^q w$ *sdescribes* $(\sigma_\gamma, aw\$, \Pi)$

iff $\gamma = \theta(q_0, \gamma)$ and there is a string z in Σ^* such that

$S \Rightarrow^*_{rm} \gamma a^q w \Rightarrow^* z$, and $(q_0, z\$, \varepsilon) \vdash^* (\sigma_\gamma, aw\$, \Pi)$.

(2) Let π^q be a nonterminal in N_p, and $\gamma\pi^q w$ a right sentential form of G. Then,

$\gamma a^q w$ *rdescribes* $(\sigma_\gamma, w\$, \Pi)$.

iff $\gamma = \theta(q_0, \gamma)$ and there is a string z in Σ^* such that

$S \Rightarrow^*_{rm} \gamma a^q w \Rightarrow^* z$, and $(q_0, z\$, \varepsilon) \vdash^* (\sigma_\gamma, w\$, \Pi)$.

Definition 4.2. Two relations *sdescribed_by* (stands for "a *shift* move is described by") and *rdescribed_by* (stands for "a *reduce* move is described by") from acceptable configurations of $LRM_k(G)$ to vocabulary strings of G are defined as follows:

(1) Let $(\sigma_\gamma, aw\$, \Pi)$ be an acceptable configuration of $LRM_k(G)$. Then,

$(\sigma_\gamma, aw\$, \Pi)$ *sdescribed_by* $\gamma a^q w$

iff $\gamma = \theta(q_0, \gamma)$, $q = \text{top}(\sigma_\gamma)$, and $ACTION(q, PREP_k(aw\$))$ contains *shift*.

(2) Let $(\sigma_\gamma, w\$, \Pi)$ be an acceptable configuration of $LRM_k(G)$. Then,

$(\sigma_\gamma, w\$, \Pi)$ *rdescribed_by* $\gamma\pi^q w$

iff $\gamma = \theta(q_0, \gamma)$, $\text{q} = \text{top}(\sigma_\gamma)$, and $ACTION(q, PREP_k(w\$))$ contains *reduce* π.

The relationship between right sentential forms of G and acceptable configurations of the LR(k) machine is shown in the following two lemmas (Lemma 4.1, and Lemma 4.2).

Lemma 4.1.

(1) *Let $\gamma a^q w$ be a right sentential form of G. Then there exists C, an acceptable configuration of $LRM_k(G)$, such that $\gamma\pi^q w$ rdescribes C.*

(2) *Let $\gamma\pi^q w$ be a right sentential form of G. Then there exists C, an acceptable configuration of $LRM_k(G)$, such that $\gamma\pi^q w$ rdescribes C.*

Proof. First, suppose that $\gamma a^q w$ is a right sentential form of G. Then there is a string z in Σ^* such that

$S \Rightarrow^*_{rm} \gamma a^q w \Rightarrow_{rm} \gamma a w \Rightarrow^* z$.

Then from Theorem 3.1, we know that there exists an LR(k) parsing sequence over G such that

$(q_0, z\$, \varepsilon) \vdash^* (\sigma_{h(\gamma)}, aw\$, \Pi) \vdash^* (\varepsilon, \$, \Pi')$

for some $\Pi, \Pi' \in P^*$.

Therefore, the relation *sdescribes* holds between $\gamma a^q w$ and $(\sigma_{h(\gamma)}, aw\$, \Pi)$.

Second, suppose that $\gamma\pi^q w$ is a right sentential form of G. Then there is a string z in Σ^* such that

$S \Rightarrow^*_{rm} \gamma\pi^q w \Rightarrow_{rm} \gamma w \Rightarrow^* z$.

Again from Theorem 3.1, we know that there exists an LR(k) parsing sequence over G such that

$(q_0,z\$,\varepsilon) \vdash^* (\sigma_{h(\gamma)},w\$,\Pi) \vdash^* (\varepsilon,\$,\Pi')$ for some $\Pi,\Pi' \in P^*$.

Now, the relation *rdescribes* holds between $\gamma\pi^q w$ and $(\sigma_{h(\gamma)},aw\$,\Pi)$.□

Lemma 4.2. *Let $(\sigma_\gamma,w\$,\Pi)$ be an acceptable configuration of $LRM_k(G)$. Then there exists α, a right sentential form of G, such that $(\sigma_\gamma,w\$^k,\Pi)$ sdescribed_by α, or $(\sigma_\gamma, w\$^k,\Pi)$ rdescribed_by α.*

Proof. First, suppose that $ACTION(q, PREP_k(w))$ contains *shift*, and top(σ_γ) is the state q.

Then there is string z in Σ^* such that

$(q_0,z\$,\varepsilon) \vdash^* (\sigma_\gamma,ay\$^k,\Pi) \vdash_{shift} (\sigma_{\gamma a},y\$^k,\Pi) \vdash^* (\varepsilon,\$^k,\Pi')$,

where $ay = w$. Then by Theorem 3.1, there exists a derivation in G such that

$S \Rightarrow^*_{rm} \theta(q_0,\gamma a)y \Rightarrow^* z$.

Since $\theta(q_0,\gamma a) = \theta(q_0,\gamma)a^q$, the relation *sdescribed_by* holds between $(\sigma_\gamma,w\$,\Pi)$ and $\theta(q_0,\gamma)a^q y$.

Second, suppose that $ACTION(q, PREP_k(w))$ contains *reduce* π with π being $A \to \alpha$, and $top(\sigma_\gamma)$ is again the state q. Then there is string z in Σ^* such that

$(q_0,z\$,\varepsilon) \vdash^* (\sigma_{\beta\alpha},w\$,\Pi) \vdash_\pi (\sigma_{\beta A},w\$,\Pi\pi) \vdash^* (\varepsilon,\$,\Pi')$,

where $\beta\alpha = \gamma$. Again from Theorem 3.1, we know that there is a derivation in G such that

$S \Rightarrow^*_{rm} \theta(q_0,\beta A)w \Rightarrow^* z$.

Since $\theta(q_0,\beta A) = \theta(q_0,\beta)A^p$ with $p = GOTO(q_0,\beta)$, Construction 3.1 says that P contains the production $A^p \to \theta(p, \alpha)\pi^{p'}$ such that $p' = GOTO(p, \alpha)$ $(= GOTO(q_0,\beta\alpha) = q)$. Thus, there exists a derivation in G such that

$S \Rightarrow^*_{rm} \theta(q_0,\beta)A^p\ w \Rightarrow_{rm} \theta(q_0,\beta)\theta(p, \alpha)\pi^p\ w \Rightarrow^* z$.

Since $\theta(q_0,\beta) \cdot \theta(p, \alpha) = \theta(q_0,\gamma)$, the relation *rdescribed_by* holds between $(\sigma_\gamma,w\$,\Pi)$ and $\theta(q_0,\gamma)\pi^p$ w.□

The following two properties (Propety 4.1 and 4.2) and two lemmas (Lemma 4.3 and 4.4) are devoted to exhibit the relationship between right sentential forms of G and moves of the LR(k) machine. Property 4.1 and 4.2 are obvious from the arguments in the proof of Lemma 4.2.

Property 4.1. *Let C and C′ be acceptable configurations of $LRM_k(G)$ such that $C \vdash_{shift} C'$. Then G has a right sentential form α such that C sdescribed_by α.*

Property 4.2. *Let C and C′ be acceptable configurations of $LRM_k(G)$ such that $C \vdash_\pi C'$. Then G has a right sentential form α such that C rdescribed_by α.*

Lemma 4.3. *Let a^q be a nonterminal in N_Σ, and $\gamma a^q\ w$ be a right sentential form of G. Then for each C such that $\gamma a^q\ w$ sdescribes C, there exists an acceptable configuration C′ such that $C \vdash_{shift} C'$.*

Proof. Since $\gamma a^q\ w$ is a right sentential form of G, there is a derivation in G such that

$S \Rightarrow^*_{rm} \delta A^p\ y \Rightarrow_{rm} \delta\alpha a^q\ \beta\pi^r\ y \Rightarrow^*_{rm} \delta\alpha a^q\ w$,

where π is $A \to h(\alpha a^q\ \beta) \in P$, $\delta\alpha = \gamma$., $p = GOTO\ (q_0,h(\delta))$, $q = GOTO\ (q_0,h(\gamma)) = GOTO\ (p,h(\alpha))$, $r = GOTO\ (q,ah(\beta))$, and

$\beta\pi' y \Rightarrow^* w$. Them according to Property 3.1-(3), there is a derivation in G such that

$S' \Rightarrow^*_{rm} h(\delta)Ay\$ \Rightarrow_{rm} h(\delta)h(\alpha)ah(\beta)y\$ \Rightarrow^*_{rm} h(\gamma)aw\$$.

From Theorem 2.2, we know that the state q, i.e., $GOTO(q_0,h(\gamma))$, contains all valid LR (k) items for the viable prefix $h(\gamma)$. Therefore the state q contains an LR(k) item $[A \rightarrow h(\alpha).ah(\beta),u]$ with $PREP_k(aw\$) \in FIRST_k(ah(\beta)u)$. According to Definition 4.1-(1), every configuration satisfying the condition of this lemma is of the form $(\sigma_{h(\gamma)},aw\$,\Pi)$. The state $top(\sigma_{h(\gamma)})$ is q because of $q=GOTO(q_0,h(\gamma))$. Therefore, $ACTION(q, PREF_k(aw\$))$ contains *shift*; and thus

$(\sigma_{h(\gamma)},aw\$,\Pi) \vdash_{shif} (\sigma_{h(\gamma)a},w\$,\Pi)$.

Moreover, Theorem 2.3 says that the configuration $(\sigma_{h(\gamma)a},w\$,\Pi)$ is acceptable because there is a string z in Σ^* such that $S' \Rightarrow^*_{rm} h(\gamma)aw\$ \Rightarrow^* z\$$, and $h(\gamma)a$ is a viable prefix of G. □

Lemma 4.4. *Let π^q be a nonterminal in N_P, and $\gamma\pi^q w$ be a right sentential form of* **G**. *Then for each C such that $\gamma\pi^q w$ rdescribes C, there exists an acceptable configuration* C′ such that $C \vdash_\pi C'$.

Proof. Assume π is $A \rightarrow \alpha$. Because the state q is the state $GOTO(q_0,h(\gamma))$ (from Property 3.1-(4)), the state q contains an LR (k) item $[A \rightarrow \alpha.,PREP_k(w\$)]$ by the similar arguments in the proof of Lemma 4.3. According to Definition 4.1-(2), every configuration satisfying the condition of this lemma, is of the form $(\sigma_{h(\gamma)},w\$,\Pi)$. Because $ACTION(q, PREP_k(w\$))$ contains *reduce* π and the state q is top $(\sigma_{h(\gamma)})$, the following move holds:

$(\sigma_{\beta\alpha},w\$,\Pi) \vdash_\pi (\sigma_{\beta A},w\$,\Pi\pi)$,

where $\beta\alpha=h(\gamma)$. Further, Theorem 2.3 says that the configuration $(\sigma_{\beta A},w\$,\Pi\pi)$ is acceptable because there is a string z in Σ^* such that $S' \Rightarrow^*_{rm} \beta Aw\$ \Rightarrow_{rm} \beta\alpha w\$ \Rightarrow^* z\$$, and βA is a viable prefix of G.□

Now Theorem 4.1 is established by Lemma 4.3 and Property 4.1, and Theorem 4.2 is established by Lemma 4.4 and Property 4.2. In virtue of the descriptive power of **G** represented by the two theorems, we present an LR(k) *machine description grammar* and our main theorem on the description of LR(k) machine as a concluding result of this paper:

Definition 4.3. Let **G** be the LR(k)-colored grammar for a CFG G. The *LR(k) machine description grammar* for G is defined by $\boldsymbol{G}_d=(S \cup N_N, N_\Sigma \cup N_p, \boldsymbol{P}_d, S)$ with

$\boldsymbol{P}_d=\mathrm{P}-(\{a^q \rightarrow a \mid a^q \in N_\Sigma\} \cup \{\pi^q \rightarrow \varepsilon \mid \pi^q \in N_p\})$.

Theorem 4.3. *Let a description sentence of z be a string α such that $\alpha \Rightarrow^* z$ and $\alpha \in$ L (G_d). Then, there is a valid LR(k) parsing sequence of z if and only if there is a decription sentence of z. (In other words, the shift and/or reduce moves in an LR(k) parsing sequence can be described by the a^q and/or π^q symbols in a description sentence of z.)*

Proof. Immediate from Definition 4.3, Theorem 4.1 and 4.2.□

REFERENCES

1. Aho, A. V. and Ullman, J.D., "The Theory of Parsing, Translation and Compiling," Vol.1-2, Prentice-Hall, Englewood Cliffs, N.J. 1972, 1973 MIT, Cambridge, Mass., 1969.
2. Aho, A. V., and Ullman, J.D., Optimization of LR(k) parsers, *J. Comput. Syst. Sci.* **6**(1972), 573-602.
3. DeRemer, F.L., Simple LR(k) grammars, *Comm. ACM* **14**(7), 453-460(1971).
4. Geller, M. M. and Harrison, M. A., On LR (k) grammars and languages, *Theoret. Comput. Sci.* **4**(1977), 245-276.
5. Heilbrunner, S., A parsing automata approach to LR theory, *Theoret. Comput. Sci.* **15**(1981), 117-157.
6. Knuth, D. E., On the translation of languages from left to right, *Information and Control* **8**, 607-639 (1965) Korea Advanced Institute of Science and Technology, 1991.
7. Lee, M. J. and Choe, K. M., SLR(k) covering for LR(k) grammars, *Information Processing Letters* **37**, 337-347(1991).
8. Tomita, M., An efficient augmented-context-free parsing algorithm, *Computational Linguistics* **13**(1-2), 31-46(1987).