Theoretical Computer Science 66 (1989) 65–85
North-Holland

ON $LC(0)$ GRAMMARS AND LANGUAGES

Rieks op den AKKER

Department of Computer Science, Twente University, P.O. Box 217, 7500 AE Enschede, The Netherlands

Communicated by M. Harrison
Received March 1987
Revised November 1987

Abstract. Several definitions of the $LR(k)$ grammars can be found in the literature. Since the left-corner grammars can be defined as a restricted class of $LR(k)$ grammars, there are also several definitions of the $LC(k)$ grammars. Two such definitions are compared. For the case $k=0$, these definitions are not equivalent. A characterization of the $LC(0)$ languages is given in terms of the simple deterministic languages and these classes of languages are compared with other classes of languages, such as the $LL(1)$ languages and the $LR(0)$ languages.

1. Introduction

Deterministic left-corner grammars or $LC(k)$ grammars were formally defined by Rosenkrantz and Lewis II [8]. These grammars are deterministically parsable by a left-corner parsing strategy. In this strategy the productions applied at a node in a derivation tree are recognized after the recognition of the *left-corner* of the production, that is the left-most symbol of the right-hand side of the production. The original definition of $LC(k)$ grammars is given in terms of left-most derivations. This definition can also be found in Aho and Ullman [1].

Soisalon-Soininen and Ukkonen have defined $LC(k)$ grammars as a restricted class of $LR(k)$ grammars [9, 10, 11]. Since there are several definitions of the $LR(k)$ grammars, there are also several possible definitions of the $LC(k)$ grammars. Geller and Harrison have given a survey of a number of different $LR(k)$ definitions [4]. Special attention is paid to the case $k=0$. In this case the several variants of the $LR(k)$ definition discussed by Geller and Harrison differ. We consider two versions of the $LC(k)$ definition, one derived from the $LR(k)$ definition proposed by Geller and Harrison [4] and one derived from the "augmented" $LR(k)$ definition of Aho and Ullman [1]. We will give a characterization of the classes of $LC(0)$ languages in terms of simple deterministic languages [7].

It can be shown [2] that a slight modification of the original definition of $LC(k)$ grammars given in [8] is equivalent with the definition of $LC(k)$ grammars from [9, 10, 11].

This paper is organized as follows. The remainder of this section is devoted to some preliminary definitions used in the other sections. In Section 2 we give the definitions of the $LR(k)$ grammars we consider in this paper and review the relevant results from [4]. In Section 3 we define the $LC(k)$ grammars derived from the $LR(k)$ grammars and we study the relation between the two different definitions of $LC(k)$ grammars. In Section 4 we consider left-recursion in $LR(0)$ grammars. Results obtained here will be used in Section 5, where we give our main result: a characterization of the $LC(0)$ languages in terms of the simple deterministic languages of Korenjak and Hopcroft [7]. Furthermore we consider the relations between the class of $LC(0)$ languages, the class of $LL(1)$ languages, and the class of $LR(0)$ languages.

The notation we use for concepts of formal language theory is—unless otherwise stated—like that in Harrison [5]. Context-free grammars are denoted by a four-tuple (N, Σ, P, S), where N and Σ are the set of nonterminal symbols and the set of terminal symbols respectively. V will denote the set $N \cup \Sigma$.

The empty string is denoted by ε. In derivations we use $\Rightarrow_r$ ($\Rightarrow_l$) to indicate that the derivation is right-most (left-most). Let p denote a production in P and $\alpha, \beta \in V^*$; then $\alpha \Rightarrow_l^p \beta$ denotes the one-step left-most derivation in which production p is applied. If $\pi = p_1 p_2 \ldots p_n$, where the p_i $(1 \leqslant i \leqslant n)$ indicate productions in P, then $\alpha \Rightarrow_l^\pi \beta$ denotes the derivation

$$\alpha \Rightarrow_l^{p_1} \alpha_1 \Rightarrow_l^{p_2} \alpha_2 \Rightarrow \cdots \Rightarrow_l^{p_n} \beta, \quad \text{where } \alpha_i \in V^*.$$

A production of the form $A \to \varepsilon$ is called an *ε-production.* A production with left-hand side A is called an *A-production.*

A context-free grammar (cfg) G is called *ε-free* if either G has no ε-productions or $S \to \varepsilon$ is the only ε-production and S does not occur in the right-hand side of any production of G. A cfg $G = (N, \Sigma, P, S)$ is said to be in *Greibach normal form* if each rule is of one of the forms

$$A \to aB_1 \ldots B_n, \qquad A \to a, \qquad S \to \varepsilon,$$

where $B_1, \ldots, B_n \in N - \{S\}$, $a \in \Sigma$.

The *length* of the string α is denoted by $|\alpha|$. For any nonnegative integer k, if $k < |\alpha|$ then $k{:}\alpha$ denotes the *prefix* of α of length k. If $k \geqslant |\alpha|$ then $k{:}\alpha$ equals α.

With respect to a cfg $G = (N, \Sigma, P, S)$, if $\alpha \in V^*$, then $L(\alpha)$ denotes the set (language) $\{w \in \Sigma^* \mid \alpha \Rightarrow^* w\}$. The *language* defined by G, denoted by $L(G)$, is the set $L(S) = \{w \in \Sigma^* \mid S \Rightarrow^* w\}$.

We call a language $L \subset \Sigma^*$ *degenerate* if $L = \emptyset$ or $L = \{\varepsilon\}$. A context-free grammar is *unambiguous* if for all $x \in L(G)$, there is exactly one left-most derivation in G.

A nonterminal $A \in N$ is said to be *left-recursive* if there exists $\alpha \in V^*$ such that $A \Rightarrow^+ A\alpha$. A context-free grammar G is said to be *left-recursive* if there exists a left-recursive nonterminal in N.

A cfg is said to be *reduced* iff for all X in V there is at least one derivation $S \Rightarrow^* \alpha X \beta \Rightarrow^* w$ for some α, β in V^* and w in Σ^*. All context-free grammars we will consider are reduced.

We will refer to the $LL(1)$ grammars and to the simple deterministic grammars of Korenjak and Hopcroft [7]. For convenience we recall the definitions of these grammars.

Definition 1.1. Let $k \geq 0$ and $G = (N, \Sigma, P, S)$ be a cfg. G is *$LL(k)$-grammar* if, for each $A \in N$, $\alpha, \beta, \gamma \in V^*$, $w, x, y \in \Sigma^*$, if $A \to \beta$ and $A \to \gamma$ are in P and

(i) $S \Rightarrow_l^* wA\alpha \Rightarrow_l w\beta\alpha \Rightarrow_l^* wx$,

(ii) $S \Rightarrow_l^* wA\alpha \Rightarrow_l w\gamma\alpha \Rightarrow_l^* wy$,

(iii) $k{:}x = k{:}y$,

then $\beta = \gamma$.

Definition 1.2. A cfg $G = (N, \Sigma, P, S)$ is a *simple deterministic grammar* if it is in Greibach normal form and for each $A \in N$, $a \in \Sigma$ and $\alpha, \beta \in V^*$, if $A \to a\alpha$ and $A \to a\beta$ are in P, then $\alpha = \beta$. Moreover, if $S \to \varepsilon$ is a production of G then this is the only production of G.

The simple deterministic grammars form a proper subclass of the $LL(1)$ grammars. The languages generated by simple deterministic grammars are the simple deterministic languages. Notice that according to Definition 1.2 the language $\{\varepsilon\}$ is a simple deterministic language, though it is not an s-language in the sense of [7]. Languages generated by $LL(1)$ grammars without ε-productions are simple deterministic languages.

2. $LR(k)$ grammars

In [4], Geller and Harrison have given an overview of the many definitions of $LR(k)$ grammars that can be found in the literature. Their definition of $LR(k)$ grammars is compared with other definitions of these grammars. Especially for $k = 0$, the definitions of these grammars are not equivalent. In this section, we give the $LR(k)$-definition of Geller and Harrison and the $LR(k)$-definition of Aho and Ullman [1]. This last class of grammars we will call the *augmented $LR(k)$* or *A-$LR(k)$* grammars. The term "$LR(k)$ grammars" will denote the $LR(k)$ grammars according to the definition of Geller and Harrison. In this section we recall the results from [4] that are relevant for our study of the $LC(k)$ grammars and languages. We start with the definition of $LR(k)$ grammars from Geller and Harrison [4].

Definition 2.1. Let $k \geq 0$ and $G = (N, \Sigma, P, S)$ be a context-free grammar such that $S \Rightarrow_r^+ S$ is impossible in G. G is $LR(k)$, if the conditions

(i) $S \Rightarrow_r^* \alpha Aw \Rightarrow_r \alpha\beta w = \gamma w$,

(ii) $S \Rightarrow_r^* \alpha' A' x \Rightarrow_r \alpha'\beta' x = \gamma w'$,

(iii) $k{:}w = k{:}w'$,

always imply that $A \to \beta = A' \to \beta'$ and $|\alpha\beta| = |\alpha'\beta'|$.

A production $A \to \beta$ of G *satisfies the* $LR(k)$ *condition* if for that particular production the conditions (i), (ii) and (iii) always imply that $A \to \beta = A' \to \beta'$ and $|\alpha\beta| = |\alpha'\beta'|$.

We now give the definition of $LR(k)$ grammars from Aho and Ullman [1]. We call these grammars A-$LR(k)$ grammars.

Definition 2.2. Let $k \geq 0$ and $G = (N, \Sigma, P, S)$ be a context-free grammar. The *augmented grammar for* G is $G' = (N', \Sigma, P', S')$, where $N' = N \cup S'$ and $P' = P \cup \{S' \to S\}$, and where S', a symbol not in $V = N \cup \Sigma$, is the new start symbol.

Definition 2.3. G is said to be A-$LR(k)$ (*augmented* $LR(k)$) if, in the augmented grammar G' for G, the conditions

(i) $S' \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha\beta w = \gamma w$,

(ii) $S' \Rightarrow_r^* \alpha' A' x \Rightarrow_r \alpha'\beta' x = \gamma w'$,

(iii) $k{:}w = k{:}w'$,

always imply that $\alpha A = \alpha' A'$ and $x = w'$.

A production $A \to \beta$ of G *satisfies the* A-$LR(k)$ *condition* if for that particular production the conditions (i), (ii) and (iii) always imply that $\alpha A = \alpha' A'$ and $x = w'$.

Notice that the consequence of Definition 2.1 is equivalent with the consequence $\alpha A = \alpha' A'$ and $x = w'$ of Definition 2.3.

$LR(k)$ grammars are unambiguous ($k \geq 0$). For a proof we refer to [5, Chapter 13]. In Geller and Harrison [4] it is shown that the classes of $LR(k)$ and A-$LR(k)$ grammars are co-extensive for all $k \geq 1$. However the definition of A-$LR(0)$ grammars is more restrictive than the $LR(0)$ definition. The following characterization of the A-$LR(0)$ grammars is given in [4].

Theorem 2.4. *Let* $G = (N, \Sigma, P, S)$ *be a context-free grammar.* G *is an* A-$LR(0)$ *grammar if and only if* G *is an* $LR(0)$ *grammar and* $S \Rightarrow_r^+ Sw$ *is impossible in* G *for any* $w \in \Sigma^+$.

The context-free grammar G_1 given by the productions $S \to Sa$ and $S \to a$, is an $LR(0)$ grammar that is not A-$LR(0)$.

In [4] it is also shown that the A-$LR(0)$ languages are the strict deterministic languages (i.e. the prefix-free deterministic context-free languages [5]). Since the grammar G_1 above generates the language a^+, which is not prefix-free, we know that the class of A-$LR(0)$ languages is properly contained in the class of $LR(0)$ languages.

In [4, 5] an $LR(0)$ language characterization theorem is given. In this theorem, a string characterization, a machine characterization and a set-theoretic characterization of the class of $LR(0)$ languages is given. Since we only use the string characterization and the set-theoretic characterization, we only give these in the following theorem.

Theorem 2.5. *Let $L \subseteq \Sigma^*$. The following three statements are equivalent.*

(a) *L is an $LR(0)$ language.*

(b) *L is a deterministic context-free language and for all $x \in \Sigma^+$, $w,y \in \Sigma^*$, if $w \in L$, $wx \in L$ and $y \in L$, then $yx \in L$.*

(c) *There exist strict deterministic languages L_0 and L_1, such that $L = L_0 L_1^*$.*

3. $LC(k)$ grammars

Soisalon-Soininen and Ukkonen have defined $LC(k)$ grammars in terms of right-most derivations as a restricted class of A-$LR(k)$ grammars [9, 10]. In fact they define the class of predictive $LR(k)$ grammars or $PLR(k)$ grammars. The $LC(k)$ grammars are properly contained in the class of $PLR(k)$ grammars. They give a transformation that transforms a $PLR(k)$ grammar, $k > 0$, into an $LL(k)$ grammar for the same language. They show that, for any integer $k > 0$, the transformed grammar is $LL(k)$ if and only if the original grammar is $PLR(k)$. The equivalence of the $LC(k)$ languages and the $LL(k)$ languages is proved in [9]. Soisalon-Soininen and Ukkonen augment the cfg with the start production $S' \to \perp S$, where $\perp$ is a new terminal symbol, instead of the production $S' \to S$. This is however only relevant for the transformation they give, not for our discussion. We now give the definition of left-corner grammars derived from the A-$LR(k)$ definition.

Definition 3.1. Let G be a cfg, $k \geqslant 0$. G is an A-$LC(k)$ grammar if each ε-production satisfies the A-$LR(k)$-condition and if in the augmented grammar G' for each production $A \to X\beta$, $X\beta \neq \varepsilon$, if

(i) $S' \Rightarrow_r^* \alpha A z_1 \Rightarrow_r \alpha X \beta z_1 \Rightarrow_r^* \alpha X y_1 z_1$,

(ii) $S' \Rightarrow_r^* \alpha' B z_2 \Rightarrow_r \alpha'\alpha'' X \gamma z_2 \Rightarrow_r^* \alpha'\alpha'' X y_2 z_2$,

(iii) $\alpha'\alpha'' = \alpha$ and $k{:}y_1 z_1 = k{:}y_2 z_2$,

then $\alpha A = \alpha' B$ and $\beta = \gamma$.

To illustrate this definition suppose that $\alpha \Rightarrow^* w$, $X \Rightarrow^* x$ and consider the terminal string wxy_1z_1 (y_1 and z_1 as in Definition 3.1). The production $A \to X\beta$ can be recognized with certainty after scanning wx and $k{:}y_1z_1$ if the grammar is A-$LC(k)$.

If we derive the left-corner grammars from the $LR(k)$ definition, we obtain the following definition of the $LC(k)$ grammars.

Definition 3.2. Let G be a cfg, $k \geqslant 0$. G is an $LC(k)$ grammar if

(1) $S \Rightarrow_r^+ S$ is impossible;

(2) each ε-production satisfies the $LR(k)$ condition; and

(3) for each production $A \to X\beta$, $X\beta \neq \varepsilon$, the conditions

(i) $S \Rightarrow_r^* \alpha A z_1 \Rightarrow_r \alpha X \beta z_1 \Rightarrow_r^* \alpha X y_1 z_1$,

(ii) $S \Rightarrow_r^* \alpha' B z_2 \Rightarrow_r \alpha'\alpha'' X \gamma z_2 \Rightarrow_r^* \alpha'\alpha'' X y_2 z_2$,

(iii) $\alpha'\alpha'' = \alpha$ and $k{:}y_1 z_1 = k{:}y_2 z_2$

imply that $\alpha A = \alpha' B$ and $\beta = \gamma$.

For example the grammar given by the productions

(1) $E \to E+T$ (2) $E \to T$

(3) $E \to T \times F$ (4) $T \to F$

(5) $F \to F \uparrow P$ (6) $F \to P$

(7) $P \to (E)$ (8) $P \to a$

is both A-$LC(1)$ and $LC(1)$.

In [10] it is shown that an A-$LC(k)$ grammar is an A-$LR(k)$ grammar. In the same way it can be shown that an $LC(k)$ grammar is an $LR(k)$ grammar. Thus $LC(k)$ grammars and A-$LC(k)$ grammars are unambiguous.

Theorem 3.3. *For $k>0$, a context-free grammar G is an A-$LC(k)$ grammar if and only if G is an $LC(k)$ grammar.*

For a proof of this theorem see the Appendix.

Theorem 3.4. *If G is an A-$LC(0)$ grammar, then G is an $LC(0)$ grammar.*

Proof. Let $G=(N, \Sigma, P, S)$ be an A-$LC(0)$ grammar. Suppose that G is not $LC(0)$. First suppose that G is not $LC(0)$ because $S \Rightarrow_r^+ S$ is possible in G. Then for some $A \in N$ and $\alpha \in V^*$, there is a production $A \to S\alpha$ in P, such that

$$S' \Rightarrow_r^* A \Rightarrow_r S\alpha \Rightarrow_r^* S \tag{3.1}$$

is a derivation in the augmented grammar G' of G. $S' \Rightarrow S$ is also a derivation in G'. It follows from this last derivation and derivation (3.1) that the productions $S' \to S$ and $A \to S\alpha$ are the same. Since $A \in N$ and $S' \notin N$ this is not possible. Thus $S \Rightarrow_r^+ S$ is not possible in G.

Suppose that G is not an $LC(0)$ grammar because there is a production $A \to \varepsilon$ in P which does not satisfy the $LR(0)$ condition. Then this ε-production does not satisfy the A-$LR(0)$ condition for the augmented grammar G'. This contradicts the assumption that G is an A-$LC(0)$ grammar.

Suppose that for some production $A \to X\beta$ ($X\beta \neq \varepsilon$) in P, the conditions (i), (ii) and (iii) of Definition 3.2 are satisfied and either $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$. Then the conditions (i), (ii) and (iii) of Definition 3.1 are also satisfied. Since $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$, it follows then that G is not an A-$LC(0)$ grammar. This however contradicts the assumption. Thus clause (3) of Definition 3.2 is also satisfied if G is an A-$LC(0)$ grammar. We finally conclude that G is an $LC(0)$ grammar. □

The inclusion of the class of A-$LC(0)$ grammars in the class of $LC(0)$ grammars is proper, since grammar G_1 in Section 2 is an $LC(0)$ grammar that is not A-$LC(0)$.

The following theorem characterizes the A-$LC(0)$ grammars in terms of the $LC(0)$ grammars. The theorem is analogous with Theorem 2.4.

Theorem 3.5. *Let $G=(N, \Sigma, P, S)$ be a context-free grammar. G is an A-$LC(0)$ grammar if and only if G is an $LC(0)$ grammar and $S \Rightarrow_r^+ Sw$ is impossible in G for any $w \in \Sigma^+$.*

Proof. For the proof of the "only-if" part of the statement, assume that $G=(N, \Sigma, P, S)$ is an A-$LC(0)$ grammar. We already have by Theorem 3.4 that G is an $LC(0)$ grammar. Since G is an A-$LR(0)$ grammar, it follows from Theorem 2.4 that $S \Rightarrow_r^+ Sw$ is impossible in G for any $w \in \Sigma^+$.

For the proof of the "if" part of the statement, assume that G is an $LC(0)$ grammar, such that $S \Rightarrow_r^+ Sw$ is impossible in G for any $w \in \Sigma^+$.

Suppose that G is not an A-$LC(0)$ grammar. It is easy to see that if G has an ε-production, which does not satisfy the A-$LR(0)$ condition, then this production does not satisfy the $LR(0)$ condition either.

Suppose that G is not an A-$LC(0)$ grammar because there is a production $A \to X\beta$ in P, which is not an ε-production, such that

$$S' \Rightarrow_r^* \alpha A z_1 \Rightarrow_r \alpha X \beta z_1 \Rightarrow_r^* \alpha X y_1 z_1 \tag{3.2}$$

and

$$S' \Rightarrow_r^* \alpha' B z_2 \Rightarrow_r \alpha'\alpha'' X \gamma z_2 \Rightarrow_r^* \alpha'\alpha'' X y_2 z_2 \tag{3.3}$$

are derivations in the augmented grammar G', where $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$, although $\alpha'\alpha'' = \alpha$.

Suppose that $X \neq S$ or $\alpha \neq \varepsilon$. Since each derivation in G' starts with production $S' \to S$, it follows from derivations (3.2) and (3.3) that the following are derivations in G.

$$S \Rightarrow_r^* \alpha A z_1 \Rightarrow_r \alpha X \beta z_1 \Rightarrow_r^* \alpha X y_1 z_1 \tag{3.4}$$

and

$$S \Rightarrow_r^* \alpha' B z_2 \Rightarrow_r \alpha'\alpha'' X \gamma z_2 \Rightarrow_r^* \alpha'\alpha'' X y_2 z_2. \tag{3.5}$$

Since $\alpha'\alpha'' = \alpha$ and $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$, it follows from derivations (3.4) and (3.5) that G is not an $LC(0)$ grammar. This, however, contradicts the assumption that G is such a grammar.

Suppose that $X = S$ and $\alpha = \varepsilon$. Then derivation (3.2) has the form

$$S' \Rightarrow_r^* A z_1 \Rightarrow_r S \beta z_1 \Rightarrow_r^* S y_1 z_1. \tag{3.6}$$

Derivation (3.6) implies that $S \Rightarrow_r^+ S y_1 z_1$ is a derivation in G. Since G is unambiguous, we have that $y_1 z_1 \in \Sigma^+$. However such a derivation was supposed to be impossible in G.

We conclude that G is an A-$LC(0)$ grammar. □

An immediate consequence of Theorems 2.4 and 3.5 is the following.

Corollary. *A context-free grammar G is an A-$LC(0)$ grammar if and only if it is an $LC(0)$ grammar and an A-$LR(0)$ grammar.*

Consider the grammars G_2 and G_3 given by the productions below:

$$\begin{array}{llll} G_2\colon & S\to bA & G_3\colon & S\to Sa\mid a \\ & S\to bB & & S\to bA\mid bB \\ & A\to a & & A\to c \\ & B\to b & & B\to b. \end{array}$$

G_2 is an A-$LR(0)$ grammar, which is not $LC(0)$. G_3 is an $LR(0)$ grammar, which is not $LC(0)$ and not A-$LR(0)$. Grammar G_1, given by the productions $S\to Sa$ and $S\to a$ is an $LC(0)$ grammar, which is not A-$LC(0)$.

4. Left-recursion

Recall that a cfg $G=(N, \Sigma, P, S)$ is left-recursive if for some $A\in N$ and some $\alpha\in V^*$, $A\Rightarrow^+ A\alpha$ is a derivation in G. In general, the existence of such a derivation in G does not imply that for some $w\in\Sigma^*$, $A\Rightarrow_r^+ Aw$ is a derivation in G.

In this section, we consider left-recursion in subclasses of the $LR(0)$ grammars. $LR(0)$ grammars may be left-recursive. Reduced strict deterministic grammars, which are A-$LR(0)$ grammars [4], are not left-recursive [5]. What can we say about left-recursion in A-$LC(0)$ grammars and $LC(0)$ grammars?

Results obtained in this section will be used in the following section where we give a characterization of the $LC(0)$ languages.

The following concept is useful. (Recall that all our context-free grammars are reduced.)

Definition 4.1. Let $G=(N, \Sigma, P, S)$ be a context-free grammar ($V=N\cup\Sigma$). Let n be an integer ($n\geq 1$). Let $A_0, A_1, \ldots, A_n\in N$ and $\alpha_1, \ldots, \alpha_n\in V^*$. If $A\in N$, then an *A-cycle* is a derivation $A\Rightarrow_l^\pi A\alpha$ in G, where π is a sequence $p_1p_2\ldots p_n$ of distinct productions p_i, where for all $1\leq i\leq n$, p_i indicates the production $A_{i-1}\to A_i\alpha_i\in P$, such that $A=A_0=A_n$. We say that the nonterminals A_i are *on the A-cycle.*

We call a context-free grammar $G=(N, \Sigma, P, S)$ *strict left-recursive* if G is left-recursive and for all left-recursive $A\in N$ if $A\Rightarrow_l^+ A\alpha$ ($\alpha\in V^*$) is a derivation in G, then this derivation has the form

$$A\Rightarrow_l^\pi A\alpha'\Rightarrow_l^* A\alpha\quad(\alpha'\in V^*),$$

where $A\Rightarrow_l^\pi A\alpha'$ is an A-cycle.

A strict left-recursive grammar $G=(N,\Sigma\ P,S)$ has one or more A-cycles (for some $A\in N$). For some nonterminal symbol A, a context-free grammar may have more than one A-cycle. A nonterminal B may be more than once on the same A-cycle. For example A is at least twice on any A-cycle.

By definition a strict left-recursive grammar is left-recursive. On the other hand, there are left-recursive grammars, which are not strict left-recursive. For example the grammar H given by the productions $S\to BSa$, $S\to b$ and $B\to\varepsilon$, is left-recursive (since $S\Rightarrow^{+} Sa$ is possible in H), although it is not strict left-recursive.

Theorem 4.2. *If a context-free grammar $G=(N,\Sigma,P,S)$ is $LR(0)$, then G is left-recursive if and only if G is strict left-recursive.*

Proof. It is by definition that a strict left-recursive grammar is left-recursive.

Let $G=(N,\Sigma,P,S)$ be a reduced $LR(0)$ grammar. Suppose that G is not strict left-recursive, although G is left-recursive. Then for some left-recursive nonterminal $A\in N$ and some integer $p\geq 1$, there are $A_0,A_1,\ldots,A_p\in N$, with $A_0=A_p=A$, $\gamma_i\in V^*$, $\delta_i\in N^*$, not all equal ε, such that for all i $(1\leq i\leq p)$ $\delta_i\Rightarrow^*\varepsilon$ and productions $A_{i-1}\to\delta_iA_i\gamma_i$. Let n_i be the number of steps in the (unique) derivation $\delta_i\Rightarrow_r^*\varepsilon$ $(1\leq i\leq p)$. Let $n=n_1+n_2+\cdots+n_p$. Notice that $n>0$, although some of the n_i may be zero.

Since G is unambiguous there exists $u\in\Sigma^+$, such that $\gamma_p\ldots\gamma_1\Rightarrow_r^* u$ and

$$A\Rightarrow_r^*\delta_1\ldots\delta_pAu \tag{$*$}$$

is a derivation in G. From this derivation we will derive a contradiction in the following way. Since G is reduced, for some $x\in\Sigma^*$, $A\Rightarrow_r^* x$ is a derivation in G. Suppose that this derivation has the following m steps.

$$A=\omega_m\Rightarrow\omega_{m-1}\Rightarrow\omega_{m-2}\Rightarrow\cdots\Rightarrow\omega_1=x, \tag{$**$}$$

where $\omega_j\in V^*$ $(1\leq j\leq m)$. Now, for some $\alpha\in V^*$ and $w\in\Sigma^*$,

$$S\Rightarrow_r^*\alpha Aw\Rightarrow_r^m\alpha xw \tag{4.1}$$

is a derivation in G, of which the last m steps are those of ($**$). Let $l\geq 1$ be an integer such that $l\times n\geq m$. Let $t=l\times n$. t is the number of steps in the (unique) derivation of ε from the sequence $(\delta_1\delta_2\ldots\delta_p)^l$. Let this derivation have the form

$$(\delta_1\ldots\delta_p)^l=\omega'_t\Rightarrow_r\omega'_{t-1}\Rightarrow_r\cdots\Rightarrow_r\omega'_1=\varepsilon, \tag{$***$}$$

where $\omega'_j\in V^*$ $(1\leq j\leq t)$. The sequence $(\delta_1\ldots\delta_p)^l$ is generated in the following derivation, in which derivation ($*$) is repeated l times.

$$\begin{aligned} S&\Rightarrow_r^*\alpha Aw\Rightarrow_r^*\alpha\delta_1\cdots\Rightarrow_r^*\delta_pAuw\Rightarrow_r^*\alpha(\delta_1\ldots\delta_p)^2Au^2w\\ &\Rightarrow_r^*\cdots\Rightarrow_r^*\alpha(\delta_1\ldots\delta_p)^lAu^lw\Rightarrow_r^*\alpha(\delta_1\ldots\delta_p)^lxu^lw\Rightarrow_r^l\alpha xu^lw. \end{aligned} \tag{4.2}$$

The last t steps in this derivation are those of ($***$).

Claim. *If $1\leq i\leq m$, then $\omega_i=\omega'_i$.*

Proof. First, we show that if $1 \leq i \leq m$, then $\omega_i = \omega'_i x_i$, for some $x_i \in \Sigma^*$. Then we will show that for all i, $x_i = \varepsilon$.

Assume, for the sake of contradiction, that for some i $(1 \leq i \leq m)$ there is no $x_i \in \Sigma^*$, such that $\omega_i = \omega'_i x_i$. Let j be the smallest integer such that there is no $x_j \in \Sigma^*$ such that $\omega_j = \omega'_j x_j$. Notice that $j > 1$, since $\omega_1 = x = \varepsilon x = \omega'_1 x$. From derivation (***) we may conclude that for all $2 \leq i \leq t$, $\omega'_i \in N^+$. Let $\omega_j = \psi B z$, $\omega_{j-1} = \psi\beta z$, $\omega'_j = \psi' B'$, $\omega'_{j-1} = \psi'\beta'$ for some $z \in \Sigma^*$, $B, B' \in N$, $\psi', \beta' \in N^*$ and $\psi, \beta \in V^*$. The production used in the right-most derivation of ω_{j-1} (ω'_{j-1}) from ω_j (ω'_j) is $B \to \beta$ $(B' \to \beta')$. Derivations

$$S \Rightarrow_r^* \alpha\omega_j w = \alpha\psi B z w \Rightarrow_r \alpha\psi\beta z w = \alpha\omega_{j-1} w$$

and

$$S \Rightarrow_r^* \alpha\omega'_j x u^l w = \alpha\psi' B' x u^l w \Rightarrow_r \alpha\psi'\beta' x u^l w = \alpha\omega'_{j-1} x u^l w$$

are derivations in G. Since $\omega_{j-1} = \omega'_{j-1} x_{j-1}$ for some $x_{j-1} \in \Sigma^*$, $\alpha\psi\beta z w = \alpha\psi'\beta' x_{j-1} w$. Since $0{:}x_{j-1} w = 0{:}x u^l w$ and since G is an $LR(0)$ grammar, we know that $\beta = \beta'$, $B = B'$ and $\psi = \psi'$. Thus $\omega_j = \psi B z = \psi' B' z = \omega'_j z$. This, however, contradicts our assumption that there is no $x_j \in \Sigma^*$ such that $\omega_j = \omega'_j x_j$. Especially $A = \omega_m = \omega'_m x_m$. This implies $A = \omega'_m$ and $x_m = \varepsilon$.

In order to show that for all i $(1 \leq i \leq m)$ $x_i = \varepsilon$, suppose that, for some i, $x_i \neq \varepsilon$. Let j be the integer $(1 \leq j < m)$ such that $x_j \neq \varepsilon$ and if $i > j$, then $x_i = \varepsilon$. Consider the derivations

$$S \Rightarrow_r^* \alpha A w \Rightarrow_r^+ \alpha\omega_{j+1} w \Rightarrow_r \alpha\omega_j w$$

and

$$S \Rightarrow_r^* \alpha A x u^l w \Rightarrow_r^+ \alpha\omega'_{j+1} x u^l w \Rightarrow_r \alpha\omega'_j x u^l w$$

in G. Since $\omega_{j+1} = \omega'_{j+1}$ and $\omega'_{j+1} \in N^+$ in the last step of these derivations the right-most symbol of ω_{j+1} is rewritten. Since $\omega_j = \omega'_j x_j$ and $0{:}x u^l w = 0{:}x_j w$, we conclude from these derivations that $\alpha\omega'_j = \alpha\omega'_j x_j$ and thus $x_j = \varepsilon$.

We conclude that, for all i, $x_i = \varepsilon$. Especially $x = \varepsilon$. □

Proof of Theorem 4.2 (*conclusion*). Since A is derivable from $\delta_1 \ldots \delta_p$ and since $A \Rightarrow^* \varepsilon$, it follows that G is ambiguous. We have derived a contradiction since G is an $LR(0)$ grammar and $LR(0)$ grammars are unambiguous. □

Theorem 4.2 can be generalized: for all integers $k \geq 0$, if a cfg G is $LR(k)$, then G is left-recursive if and only if G is strict left-recursive. This generalization can be proved in essentially the same way. Notice that this result holds for $LL(k)$ grammars, $LC(k)$ grammars and all other subclasses of the $LR(k)$ grammars.

$LC(0)$ grammars and A-$LR(0)$ grammars may be left-recursive. The grammar G_4 given by the productions

$$S \to Aa, \qquad A \to Bb, \qquad B \to Ab, \qquad A \to \varepsilon$$

is A-$LR(0)$ and not $LC(0)$. The grammar G_5 given by the productions

$$S \to Aa, \qquad A \to Bb, \qquad B \to S, \qquad A \to \varepsilon$$

is $LC(0)$ and not A-$LR(0)$.

Theorem 4.3. *Let G be an $LC(0)$-grammar.*

(a) *If G has an A-cycle for some $A \in N$, then S is on this A-cycle.*

(b) *G has at most one S-cycle.*

Proof. (a): Let n be a positive integer. Let $A_0, A_1, \ldots, A_n \in N$ and $\alpha_1, \ldots, \alpha_n \in V^*$. Let $G = (N, \Sigma, P, S)$ have the A-cycle $A \Rightarrow_l^\pi A\alpha$. Let π be the sequence $p_1 p_2 \ldots p_n$ where, for all $1 \leq i \leq n$, p_i indicates the production $A_{i-1} \to A_i\alpha_i \in P$ such that $A = A_0 = A_n$. Suppose that S is not on this A-cycle. Since G is reduced, there is an integer j $(1 \leq j \leq n)$ such that A_j occurs in the right-hand side of a production not equal to production p_j. Let this production be $B \to \gamma_1 A_j \gamma_2$, where $B \in N$ and $\gamma_1, \gamma_2 \in V^*$. Then the derivations

$$S \Rightarrow_r^* \beta B w_1 \Rightarrow_r \beta\gamma_1 A_j \gamma_2 w_1 \Rightarrow_r^* \beta\gamma_1 A_j v_1 w_1 \tag{4.3}$$

and

$$S \Rightarrow_r^* \beta\gamma_1 A_{j-1} w_2 \Rightarrow_r \beta\gamma_1 A_j \alpha_j w_2 \Rightarrow_r^* \beta\gamma_1 A_j v_2 w_2 \tag{4.4}$$

are derivations in G. The first right sentential form of derivation (4.4) results from the derivation

$$S \Rightarrow_r^* \beta\gamma_1 A_j v_1 w_1 \Rightarrow_r^* \beta\gamma_1 A_{j-1} w_2.$$

Since $0{:}v_1 w_1 = 0{:}v_2 w_2$, it follows from derivations (4.3) and (4.4) and clause (3) of Definition 3.2 that the productions $B \to \gamma_1 A_j \gamma_2$ and $A_{j-1} \to A_j \alpha_j$ are the same. This contradicts the assumption that they were not the same. We conclude that S is on the A-cycle.

In order to prove (b), suppose that G has two distinct S-cycles. It is easy to show that G does not satisfy clause (3) of Definition 3.2. □

Theorem 4.3 implies that each nonterminal (except for S) is at most once on the S-cycle of an $LC(0)$ grammar. Grammar G_4 above has an A-cycle and S is not on this A-cycle. Therefore G_4 is not an $LC(0)$ grammar. G_5 satisfies propositions (a) and (b) in Theorem 4.3.

Theorem 4.4. *An A-$LC(0)$ grammar is not left-recursive.*

Proof. Let $G = (N, \Sigma, P, S)$ be an A-$LC(0)$ grammar. From Theorem 3.5 it follows that G has no S-cycle. Since G is an $LC(0)$ grammar, we know from Theorem 4.3 that G is not strict left-recursive. From Theorem 4.2 it follows that G is not left-recursive. □

Theorem 4.3 and 4.4 cannot be generalized to $LC(k)$ grammars or A-$LC(k)$ grammars with arbitrary look-ahead. The grammar given after Definition 3.2 illustrates this.

5. $LC(0)$ languages

For positive k, we know that the $LC(k)$ languages and the A-$LC(k)$ languages are the $LL(k)$ languages. In this section we first show that the A-$LC(0)$ languages are the simple deterministic languages of Korenjak and Hopcroft [7]. Then we give a characterization of the $LC(0)$ languages in terms of simple deterministic languages and show that the $LC(0)$ languages are properly contained in the $LL(1)$ languages and properly contain the class of simple deterministic languages.

Theorem 5.1. *If G is a simple deterministic grammar, then G is an A-LC(0) grammar.*

Proof. This is an immediate consequence of the definitions of simple deterministic grammars and A-$LC(0)$ grammars. □

Corollary. *The class of simple deterministic languages is contained in the class of A-LC(0) languages.*

We now show that A-$LC(0)$ languages are simple deterministic languages. A-$LC(0)$ grammars may have ε-productions. For example, the grammar G_6 given by the productions

$$S \to A|aB, \qquad A \to b, \qquad B \to \varepsilon$$

is an A-$LC(0)$ grammar. We will show that if an A-$LC(0)$ grammar has an A-production which is an ε-production, then this is the only A-production in the grammar. Therefore we first give the following more general result.

Theorem 5.2. *Let $G = (N, \Sigma, P, S)$ be an LR(0) grammar. Let $A \in N$ and $\beta_1, \beta_2 \in V^*$. If $A \to \beta_1$ and $A \to \beta_1\beta_2$ are distinct productions in P, then* (i) $\beta_1 = \varepsilon$ *and* (ii) *for some $B \in N$ and $\delta \in V^*$, $\beta_2 = B\delta$ and* (iii) *G has an A-cycle.*

Proof. Let $A \to \beta_1$ and $A \to \beta_1\beta_2$ be productions in P, where $\beta_2 \neq \varepsilon$. For some $\alpha, \gamma \in V^*$ and some $w \in \Sigma^*$,

$$S \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha\beta_1 w = \gamma w \tag{5.1}$$

is a derivation in G. The string β_2 is either an element of Σ^+ or it contains at least one nonterminal symbol.

In the former case let $\beta_2 = u$ for some $u \in \Sigma^+$. Then the following derivation exists in G.

$$S \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha\beta_1 u w = \gamma w'. \tag{5.2}$$

Since in derivations (5.1) and (5.2) $0{:}w = 0{:}w'$, it follows from Definition 2.1 that the productions $A \to \beta_1$ and $A \to \beta_1\beta_2$ are the same. This contradicts the assumption that β_2 is not the empty string.

We now consider the case that $\beta_2 = zB\delta$, for some $z \in \Sigma^*$, $B \in N$ and $\delta \in V^*$. Since G is a reduced grammar, for some $u, v, x, y \in \Sigma^*$ and some production $D \to y$ in P,

$$S \Rightarrow_r^* \alpha Aw \Rightarrow_r \alpha\beta_1 zB\delta w \Rightarrow_r^* \alpha\beta_1 zBxw$$
$$\Rightarrow_r^* \alpha\beta_1 zvDuxw \Rightarrow_r \alpha\beta_1 zvyuxw = \gamma w'' \tag{5.3}$$

is a derivation in G. Hence the production $D \to y$ is the last production used in the right-most derivation of the terminal string vyu from B. Since in derivations (5.1) and (5.3) $0{:}w = 0{:}w''$, we must conclude from Definition 2.1 that the productions $A \to \beta_1$ and $D \to y$ are the same and that $\alpha = \alpha\beta_1 zv$. This implies that $y = \beta_1 = z = v = \varepsilon$. Hence, our two productions $A \to \beta_1$ and $A \to \beta_1\beta_2$ are of the form $A \to \varepsilon$ and $A \to B\delta$. Moreover, it follows from derivation (5.3) that there is an A-cycle in G. □

Since an A-$LC(0)$ grammar is $LR(0)$ (see Theorem 3.4) and not left-recursive (see Theorem 4.4), it follows from Theorem 5.2 that if $A \to \varepsilon$ is a production in the grammar then this is the only A-production. This implies that if $S \to \varepsilon$ is a production of an A-$LC(0)$ grammar then this is the only production of the grammar and thus S does not occur in the right-hand side of a production. It can be shown that if $G = (N, \Sigma, P, S)$ is an A-$LC(0)$ grammar, then for all $A \in N$, if $\varepsilon \in L(A)$, and $L(A) = \{\varepsilon\}$.

Theorem 5.3. *Each A-$LC(0)$ grammar is equivalent to an ε-free A-$LC(0)$ grammar.*

Proof. Let $G = (N, \Sigma, P, S)$ be an A-$LC(0)$ grammar. Then G has no A-cycle for any $A \in N$ (Theorem 4.4). Let $A \to \varepsilon$ $(A \neq S)$ be a production in P. Then there is no other A-production in P (see our conclusion after the proof of Theorem 5.2). We now transform G into a new grammar in the following way.

Remove the A-production from P. Remove all occurrences of the symbol A in the right-hand side of the productions in P (i.e. substitute ε for A whenever A occurs in the right-hand side of a production). The resulting grammar $G' = (N', \Sigma, P', S)$ obviously generates the same language as G does.

Notice that by this transformation new ε-productions can be introduced. For instance, if $B \to A$ is a production in the original grammar, then the resulting grammar has the production $B \to \varepsilon$. It is easy to verify that the resulting grammar is an A-$LC(0)$ grammar. Thus we can repeat the transformation until we have an A-$LC(0)$ grammar without ε-productions. □

Theorem 5.4. (a) *If G is an ε-free A-$LC(0)$ grammar, then G is an $LL(1)$ grammar.*

(b) *If G is an ε-free $LL(1)$ grammar, which does not have the production $S \to \varepsilon$, then G is an A-$LC(0)$ grammar.*

Proof. (a): Let $G = (N, \Sigma, P, S)$ be an ε-free A-$LC(0)$ grammar. Suppose that G is not an $LL(1)$ grammar. Then there exist distinct productions $A \to \alpha_1$ and $A \to \alpha_2$ in P such that, for some $a \in \Sigma$ and some $\gamma_1, \gamma_2 \in V^*$,

$$A \Rightarrow \alpha_1 \Rightarrow_l^* a\gamma_1 \quad \text{and} \quad A \Rightarrow \alpha_2 \Rightarrow_l^* a\gamma_2$$

are derivations in G. These left-most derivations are supposed to be the shortest derivations that derive a string with a as the first symbol. Since $\alpha_1 \neq \alpha_2$ these derivations differ in at least one step. Let the distinct productions $B_1 \to X\beta_1$ and $B_2 \to X\beta_2$ in P, where $X \in N \cup \{a\}$ and $\beta_1, \beta_2 \in V^*$, be the productions used in the last of the different steps of these derivations. Then

$$S \Rightarrow_r^* aAw \Rightarrow_r^* \alpha B_1 u_1 w \Rightarrow_r \alpha X \beta_1 u_1 w \Rightarrow_r^* \alpha X z_1 u_1 w \tag{5.4}$$

and

$$S \Rightarrow_r^* \alpha Aw \Rightarrow_r^* \alpha B_2 u_2 w \Rightarrow_r \alpha X \beta_2 u_2 w \Rightarrow_r^* \alpha X z_2 u_2 w \tag{5.5}$$

are derivations in G. The B-productions may be equal to the distinguished A-productions. Since $0{:}z_1u_1w = 0{:}z_2u_2w$, it follows from derivations (5.4) and (5.5) and Definition 3.1 that G is not an $A\text{-}LC(0)$ grammar, contradicting our assumption. We conclude that G is an $LL(1)$ grammar.

(b): First, notice that the ε-free $LL(1)$ grammar with the productions $S \to \varepsilon$ and $S \to a$ is not an $A\text{-}LC(0)$ grammar. Thus the condition that G does not have production $S \to \varepsilon$ is necessary. Assume that G is an ε-free $LL(1)$ grammar and that $S \to \varepsilon$ is not a production of G. Let G' be the augmented grammar of G. Suppose that G is not an $A\text{-}LC(0)$ grammar. Since G has no ε-productions, this means that for some $A, B \in N$, $\alpha, \alpha', \alpha'', \beta, \gamma \in V^*$; $X \in V$ and $w, w' \in \Sigma^*$,

$$S' \Rightarrow_r^* \alpha Aw \Rightarrow_r \alpha X \beta w \tag{5.6}$$

and

$$S' \Rightarrow_r^* \alpha' Bw' \Rightarrow_r \alpha' \alpha'' X \gamma w' \tag{5.7}$$

are derivations in G', such that $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$, although $\alpha = \alpha' \alpha''$.

Now, let $\alpha \Rightarrow^* u$, $X \Rightarrow^* x$, $\beta \Rightarrow^* v$ and $\gamma \Rightarrow^* y$. Since G has no ε-productions, $x \in \Sigma^+$ (if $X \in \Sigma$ then $X = x$). Because of the existence of derivations (5.6) and (5.7) in G', we know that $uxvw$, $uxyw' \in L(G)$. In the leftmost derivation of $uxvw$ the production $A \to X\beta$ is used. Let this derivation have the form

$$S' \Rightarrow_l^\pi uA\omega \Rightarrow_l uX\beta\omega \Rightarrow_l^* uxvw, \tag{5.8}$$

in which π denotes a sequence of productions of G'. Since G is $LL(1)$ grammar, the left-most derivation of $uxyw'$ has the form

$$S' \Rightarrow_l^\pi uA\omega \Rightarrow_l uX\beta\omega \Rightarrow_l^* uxyw'. \tag{5.9}$$

It follows from derivation (5.7) that the production $B \to \alpha'' X\gamma$ is used in the left-most derivation of $uxyw'$. Moreover, in this left-most derivation, α'' derives a string $u_2 \in \Sigma^*$, such that $u = u_1 u_2$ for some $u_1 \in \Sigma^*$. Since we have assumed that this production is not the production $A \to X\beta$, there are two possibilities: either $B \to \alpha'' X\gamma$ is used before or after the use of production $A \to X\beta$ in (5.9). In the latter case α'' derives the empty string because u is already derived before this particular application of production $A \to X\beta$ in the left-most derivation. Since G is ε-free this implies that $\alpha'' = \varepsilon$ and we must conclude that X is a left-recursive nonterminal. This, however, contradicts the assumption that G is an $LL(1)$ grammar. If the production

$B \to \alpha'' X\gamma$ is used before the use of $A \to X\beta$ in (5.9) then it follows in the same way that X is left-recursive and again we have a contradiction. We conclude that G is an A-$LC(0)$ grammar. □

Theorem 5.5. *If L is an A-$LC(0)$ language, then L is a simple deterministic language.*

Proof. Let L be an A-$LC(0)$ language. It follows from Theorem 5.3 that L has an ε-free A-$LC(0)$ grammar G. Either $L = \{\varepsilon\}$ or $\varepsilon \notin L$. In the first case L is simple deterministic. In the second case it follows from Theorem 5.4(a) that G is an $LL(1)$ grammar without ε-productions and thus L is a simple deterministic language. □

It follows from this last theorem and the Corollary of Theorem 5.1 that the class of simple deterministic languages and the class of A-$LC(0)$ languages are the same class of languages.

We proceed with the characterization of the $LC(0)$ languages. Recall that a language $L \subset \Sigma^*$ is degenerate if $L = \emptyset$ or $L = \{\varepsilon\}$.

Lemma 5.6. *If L_1 and L_2 are simple deterministic languages, then $L = L_1 L_2^*$ is an $LC(0)$ language.*

Proof. We first consider the special case in which L_1, L_2 or both are degenerate languages. Suppose $L_1 = \emptyset$. Then $L = \emptyset$ and L is an $LC(0)$ language. Suppose $L_2 = \emptyset$ or $\{\varepsilon\}$. Then $L = L_1$ and since L_1 is a simple deterministic language, L_1 is an A-$LC(0)$ language and thus an $LC(0)$ language.

We now treat the nondegenerate case. Since L_1 and L_2 are simple deterministic languages, there exist simple deterministic grammars $G_1 = (N_1, \Sigma_1, P_1, S_1)$ and $G_2 = (N_2, \Sigma_2, P_2, S_2)$, such that $L_1 = L(G_1)$ and $L_2 = L(G_2)$, with $N_1 \cap N_2 = \emptyset$.

Define $G = (N_1 \cup N_2, \Sigma_1 \cup \Sigma_2, P_1 \cup P_2 \cup \{S_1 \to S_1 S_2\}, S_1)$. Clearly $L(G) = L_1 L_2^*$. In order to prove that G is an $LC(0)$ grammar, suppose that it is not. Since G_1 is simple deterministic and since S_2 does not derive the empty string in G_2, $S_1 \Rightarrow_r^+ S_1$ is not possible in G. Furthermore, there are no ε-productions in G, since these productions do not occur in G_1 and G_2. This means that—since G is supposed to be not an $LC(0)$ grammar—there is a production $A \to X\beta$ $(X\beta \neq \varepsilon)$ of G and

$$S_1 \Rightarrow_r^* \alpha A z_1 \Rightarrow_r \alpha X\beta z_1 \Rightarrow_r^* \alpha X y_1 z_1 \tag{5.10}$$

and

$$S_1 \Rightarrow_r^* \alpha' B z_2 \Rightarrow_r \alpha'\alpha'' X\gamma z_2 \Rightarrow_r^* \alpha'\alpha'' X y_2 z_2 \tag{5.11}$$

are derivations in G, such that $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$, although $\alpha'\alpha'' = \alpha$ and (trivially) $0{:}y_1 z_1 = 0{:}y_2 z_2$. Since the production $S_1 \to S_1 S_2$ is the only production of G in which S_1 occurs in the right-hand side (recall that G_1 and G_2 are grammars in Greibach normal form), we know that $A \to X\beta$ is not the production $S_1 \to S_1 S_2$. Since G_1 and G_2 are simple deterministic grammars, we know that $X \in \Sigma$ and that $\alpha'' = \varepsilon$. Thus $\alpha = \alpha'$. It follows from the construction of grammar G from G_1 and G_2, that if

$A \in N_2$, then $1{:}\alpha = S_1$ (recall that $1{:}\alpha$ denotes the first symbol of α). The same holds for B in derivation (5.11), i.e. $B \in N_2$ if and only if $1{:}\alpha' = S_1$. Since $\alpha = \alpha'$, either A and B both in N_2 or both in N_1. Suppose that $A, B \in N_2$. Let $\alpha = \alpha' = S_1\delta$, for some $\delta \in (\Sigma \cup N_2)^*$. It follows from the existence of the derivations (5.10) and (5.11) in G that the following two derivations exist in G_2:

$$S_2 \Rightarrow_{\mathrm{r}}^* \delta A w_1 \Rightarrow_{\mathrm{r}} \delta X \beta w_1 \Rightarrow_{\mathrm{r}}^* \delta X y_1 w_1, \tag{5.12}$$

$$S_2 \Rightarrow_{\mathrm{r}}^* \delta B w_2 \Rightarrow_{\mathrm{r}} \delta X \gamma w_2 \Rightarrow_{\mathrm{r}}^* \delta X y_2 w_2, \tag{5.13}$$

where $w_1 = u_1 z_1$ and $w_2 = u_2 z_2$ for some $u_1, u_2 \in \Sigma^*$. Since $\beta \neq \gamma$, it follows from derivations (5.12) and (5.13) that G_2 is not an $LC(0)$ grammar. It follows however by Theorem 3.4 and Theorem 5.1 and the assumption that G_2 is a simple deterministic grammar, that G_2 is an $LC(0)$ grammar. Thus we have a contradiction.

Suppose that $A, B \in N_1$. In the same way, we can construct derivations in G_1 from derivations (5.10) and (5.11) and show that G_1 is not an $LC(0)$ grammar, contradicting the assumption that G_1 is a simple deterministic grammar and an $LC(0)$ grammar. We finally conclude that G is an $LC(0)$ grammar. □

Lemma 5.7. *If L is an $LC(0)$ language then L can be written as $L_1L_2^*$, where L_1 and L_2 are simple deterministic languages.*

Proof. Let $G = (N, \Sigma, P, S)$ be a reduced $LC(0)$ grammar for L. Suppose that G has not an S-cycle. Then, by Theorem 3.5, G is an A-$LC(0)$ grammar and by Theorem 5.5, L is a simple deterministic language. Then $L = L_1L_2^*$, with $L_1 = L$ and $L_2 = \{\varepsilon\}$.

Suppose that G has an S-cycle. By Theorem 4.3, G has only one S-cycle. Let this S-cycle be $S \Rightarrow_l^{\pi_0} S\alpha$, where $\pi_0 = p_1p_2 \ldots p_n$, with $p_i \in P$ for all i, $1 \leq i \leq n$, denotes the production $A_{i-1} \to A_i\alpha_i$, $\alpha = \alpha_1\alpha_2 \ldots \alpha_n$ and $A_0 = S = A_n$.

Consider the left-most derivations of sentences of L. They have the form $S \Rightarrow_l^{\pi} z$, where $\pi = \pi_0^t\pi'$, for some integer $t \geq 0$ and some π' such that π_0 is not a prefix of π'. Let L' be the set $\{u \mid S \Rightarrow_l^{\pi'} u$, for some π' such that π_0 is not a prefix of $\pi'\}$. Clearly $L = L'(L(\alpha))^*$. We will now show that L' and $L(\alpha)$ are both simple deterministic languages.

Let P_1 be the set of productions $P - \{A_{n-1} \to A_n\alpha_n\}$. Let $G_S = (N, \Sigma, P_1, S)$. Clearly G_S is an $LC(0)$ grammar. Since $S \Rightarrow_l^{\pi_0} S\alpha$ is not a derivation in G_S and since this is the only S-cycle in G, $S \Rightarrow_{\mathrm{r}}^* Sw$ is impossible in G_S, for any $w \in \Sigma^+$. It follows from Theorem 3.5 that G_S is an A-$LC(0)$ grammar. Thus $L(G_S)$ is a simple deterministic language.

$L' = L(G_S)$. To show this, we need the following claim.

Claim A. *Let $\bar{P}$ denote the set of productions $P - \{p_1, p_2, \ldots, p_n\}$. For all i ($1 \leq i \leq n$), A_i does not occur in α and not in the right-hand side of any production in $\bar{P}$.*

Proof. Suppose that there are i, j ($1 \leq i, j \leq n$), such that A_i occurs in α_j. Then

$\alpha_j = \delta A_i \gamma$ for some $\delta, \gamma \in V^*$. Furthermore,

$$S \Rightarrow_r^* A_{j-1} w \Rightarrow_r A_j \delta A_i \gamma w \Rightarrow_r^* A_j \delta A_i z w \tag{5.14}$$

and

$$S \Rightarrow_r^* A_{j-1} w \Rightarrow_r A_j \delta A_i \gamma w \Rightarrow_r^* A_j \delta A_i z w$$
$$\Rightarrow_r^* A_j \delta A_{i-1} uzw \Rightarrow_r A_j \delta A_i \alpha_i uzw \Rightarrow_r^* A_j \delta A_i vuzw \tag{5.15}$$

are derivations in G. Since G is an $LC(0)$ grammar, it follows from derivations (5.14) and (5.15) that $A_j\delta = \varepsilon$. This is impossible. We conclude that there is no occurrence of A_i in α.

We now prove the second part of Claim A. Suppose that, for some $B \in N$, $\delta, \gamma \in V^*$ and $1 \leq j \leq n$, $B \to \delta A_j \gamma$ is a production in $\bar{P}$. Then

$$S \Rightarrow_r^* \alpha B w \Rightarrow_r \alpha \delta A_j \gamma w \Rightarrow_r^* \alpha \delta A_j z w \tag{5.16}$$

and

$$S \Rightarrow_r^* \alpha B w \Rightarrow_r \alpha \delta A_j \gamma w \Rightarrow_r^* \alpha \delta A_j z w$$
$$\Rightarrow_r^* \alpha \delta A_{j-1} uzw \Rightarrow_r \alpha \delta A_j \alpha_j uzw \Rightarrow_r^* \alpha \delta A_j vuzw \tag{5.17}$$

are derivations in G. Since G is an $LC(0)$ grammar, we conclude from derivations (5.16) and (5.17) that $A_{j-1} \to A_j \alpha_j = B \to \delta A_j \gamma$. This is impossible, since the first one is in $\bar{P}$ and the second one is not in $\bar{P}$. We conclude that the A_i do not occur in the right-hand side of a production in $\bar{P}$. □

We now show that $L' = L(G_S)$. Let $u \in L'$. From Claim A, it follows that production $A_{n-1} \to A_n \alpha_n$ is not used in the derivation of u. Thus $u \in L(G_S)$. On the other hand, let $u \in L(G_S)$. Then there is a derivation $S \Rightarrow_l^{\bar{\pi}} u$ in G, and π_0 is not a prefix of $\bar{\pi}$. Thus $u \in L'$. Since $L' = L(G_S)$ and $L(G_S)$ is a simple deterministic language, L' is a simple deterministic language.

Claim B. *For all $A \in N$, if A occurs in α, then $L_G(A)$ is a simple deterministic language.*

Proof. Let $\bar{P}$ denote the same set as in Claim A. Let $A \in N$ occur in α. Consider the context-free grammar $G_A = (N, \Sigma, \bar{P}, A)$. In the same way as is done for G_S above, it can be shown that G_A is an A-$LC(0)$ grammar. Thus $L(G_A)$ is a simple deterministic language.

We now prove that $L_G(A) = L(G_A)$. Let $u \in L_G(A)$. Suppose that for some j ($1 \leq j \leq n$) the production $A_{j-1} \to A_j \alpha_j$ is used in the derivation of u from A in G. This contradicts Claim A. Thus $u \in L(G_A)$.

On the other hand, since $\bar{P} \subseteq P$, it follows that $L(G_A) \subseteq L_G(A)$.

We conclude that $L_G(A)$ is a simple deterministic language. □

Since the simple deterministic languages are closed under product [7], it follows from Claim B, that $L(\alpha)$ is a simple deterministic language. With $L_1 = L'$ and $L_2 = L(\alpha)$, $L = L_1 L_2^*$. □

From Lemmas 5.6 and 5.7, the following characterization theorem for $LC(0)$ languages is obtained.

Theorem 5.8. *A context-free language L is an $LC(0)$ language if and only if there are simple deterministic languages L_1 and L_2, such that $L = L_1L_2^*$.*

Let $L = L_1L_2^*$ be a nonempty $LR(0)$ language, where L_1 and L_2 are strict deterministic languages. The Unique Factorization Theorem for $LR(0)$ languages (cf. [4] or [5, p. 524]) says: *If there are two strict deterministic languages L_1' and L_2' such that $L = L_1'(L_2')^*$, then $L_1 = L_1'$ and either*

(i) *$L_2 = L_2'$, or*

(ii) *L_2, L_2' are degenerate.*

Now let $L = L_1L_2^*$ be a nonempty $LC(0)$ language, where L_1 and L_2 are simple deterministic languages. Simple deterministic languages are strict deterministic [5]. Since $LC(0)$ languages are $LR(0)$ languages, the factorization of L is unique. If we read "simple" instead of "strict" in the Unique Factorization Theorem for $LR(0)$ languages, we obtain the Unique Factorization Theorem for $LC(0)$ languages.

Since $LC(0)$ grammars are $LC(1)$, and since $LC(1)$ languages are exactly the $LL(1)$ languages [9], we know that the $LC(0)$ languages are contained in the class of $LL(1)$ languages. This inclusion is proper. To see this, consider the language $L_a = a^*b^*$. $L_a = L(G_7)$, where G_7 is given by the productions

$$S \to aS, \qquad S \to bA, \qquad S \to \varepsilon, \qquad A \to bA, \qquad A \to \varepsilon.$$

Since G_7 is an $LL(1)$ grammar, L_a is an $LL(1)$ language.

We show that L_a is not an $LR(0)$-language. Therefore we use the following string characterization of $LR(0)$ languages from [4] (see also Theorem 2.5). If $L \subseteq \Sigma^*$ is an $LR(0)$-language, then for all $x \in \Sigma^+$, $w, y \in \Sigma^*$, if $w \in L$, $wx \in L$ and $y \in L$, then $yx \in L$. Suppose that L_a is an $LR(0)$-language. Since a^n, a^na^m and b^k are elements of L_a, it follows from the string characterization of $LR(0)$ languages that b^ka^m is an element of L_a. This is however not the case. Thus L_a is not an $LR(0)$-language.

The intersection of the class of $LC(0)$-languages and the class of strict deterministic languages (or prefix-free deterministic languages) is the class of simple deterministic languages. To see this, let L be a strict deterministic language, which is not simple. Suppose that L is an $LC(0)$-language. Then there exist nondegenerate simple deterministic languages L_0 and L_1, such that $L = L_0L_1^*$. This however implies that L is not a prefix-free language. (This result follows also immediately from the Corollary in Section 3.)

The relations between the classes of languages considered here are depicted in Fig. 1. $L_b = \{a^n(bd + b + c)^n\$ \mid n \geq 1\}$. This language is a prefix-free $LL(1)$ language, which is not simple deterministic (cf. [6]). $L_c = \{a^nb^n, a^nc^n \mid n \geq 1\}$. This language is a strict deterministic language which is not LL. (cf. [3] for a proof that L_c is not an LL-language) $L_d = L_b\{a\}^*$. Since L_b is a strict but not simple deterministic language, L_d is an $LR(0)$ language, which is not $LC(0)$. It is easy to verify that L_d

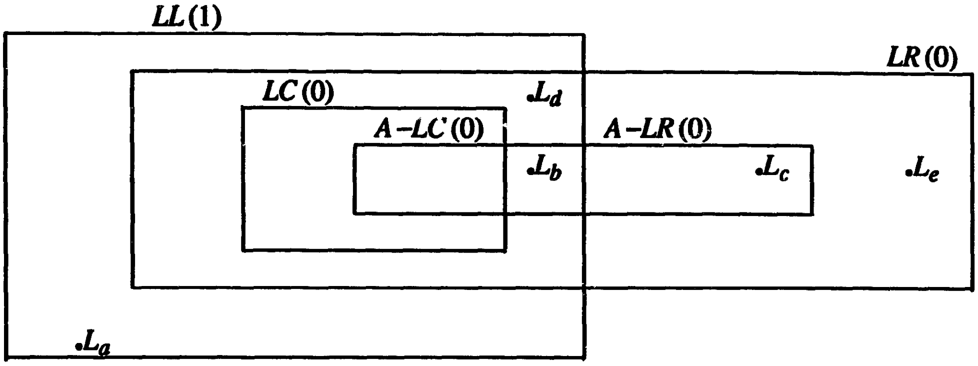

Fig. 1. Comparison of classes of languages.

is an $LL(1)$-language. $L_e = L_c\{a\}^*$. This language is $LR(0)$, not strict deterministic and not an LL-language.

Appendix

Proof of Theorem 3.3. We first prove the "only-if" part of the statement. Let $G = (N, \Sigma, P, S)$ be an A-$LC(k)$ grammar for some $k > 0$. Let G' be the augmented grammar of G. Suppose that G is not an $LC(k)$ grammar. We first assume that G is not an $LC(k)$ grammar, because $S \Rightarrow_r^+ S$ is a derivation in G. Then for some $A \in N$ and $\alpha \in V^*$, $A \to S\alpha$ is a production in G, such that

$$S' \Rightarrow_r^+ A \Rightarrow_r S\alpha \Rightarrow_r^* S$$

is a derivation in G'. Since $S' \Rightarrow_r S$ is also a derivation in G', it follows from the definition of A-$LC(k)$ grammars that the productions $A \to S\alpha$ and $S' \to S$ are the same. This is impossible and thus $S \Rightarrow_r^+ S$ is not a derivation in G.

Assume that G is not an $LC(k)$ grammar, because there is an ε-production $A \to \varepsilon$, which does not satisfy the $LR(k)$ condition. Then

$$S \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha w = \gamma w \tag{A.1}$$

and

$$S \Rightarrow_r^* \alpha' A' x \Rightarrow_r \alpha' \beta' x = \gamma w' \tag{A.2}$$

are derivations in G such that $\alpha A \neq \alpha' A'$ or $x \neq w'$, although $k{:}w = k{:}w'$. It follows from the construction of G' that

$$S' \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha w = \gamma w \tag{A.3}$$

and

$$S' \Rightarrow_r^* \alpha' A' x \Rightarrow_r \alpha' \beta' x = \gamma w' \tag{A.4}$$

are derivations in G' such that $k{:}w = k{:}w'$. Since $\alpha A \neq \alpha' A'$ or $x \neq w'$, it follows from derivations (A.3) and (A.4) that the production $A \to \varepsilon$ does not satisfy the

A-$LR(k)$ condition. Thus G is not an A-$LC(k)$ grammar. This, however, contradicts our assumption. We conclude that each ε-production of G satisfies the $LR(k)$-condition.

Finally, suppose that there is a production $A \to X\beta$, with $X\beta \neq \varepsilon$ in P, such that clause (3) of Definition 3.2 is not satisfied. Then there are derivations

$$S \Rightarrow_r^* \alpha A z_1 \Rightarrow_r \alpha X\beta z_1 \Rightarrow_r^* \alpha X y_1 z_1 \tag{A.5}$$

and

$$S \Rightarrow_r^* \alpha' B z_2 \Rightarrow_r \alpha'\alpha'' X\gamma z_2 \Rightarrow_r^* \alpha'\alpha'' X y_2 z_2 \tag{A.6}$$

in G such that $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$, although $\alpha'\alpha'' = \alpha$ and $k{:}y_1z_1 = k{:}y_2z_2$. This implies that there are derivations

$$S' \Rightarrow_r^* \alpha A z_1 \Rightarrow_r \alpha X\beta z_1 \Rightarrow_r^* \alpha X y_1 z_1 \tag{A.7}$$

and

$$S' \Rightarrow_r^* \alpha' B z_2 \Rightarrow_r \alpha'\alpha'' X\gamma z_2 \Rightarrow_r^* \alpha'\alpha'' X y_2 z_2 \tag{A.8}$$

in G' such that $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$, although $\alpha'\alpha'' = \alpha$ and $k{:}y_1z_1 = k{:}y_2z_2$. This implies that G is not an A-$LC(k)$ grammar, contradicting the assumption that G is such a grammar. We conclude that G is an $LC(k)$ grammar.

We now prove the "if" part of the statement. Let G be an $LC(k)$ grammar. First, suppose that G is not an A-$LC(k)$ grammar because there is an ε-production, which does not satisfy the A-$LR(k)$-condition. Then (A.3) and (A.4) are derivations in G' such that $\alpha A \neq \alpha' A'$ and $x \neq w'$, although $k{:}w = k{:}w'$. Since every derivation in G' starts with the production $S' \to S$, this implies that (A.1) and (A.2) are derivations in G such that $k{:}w = k{:}w'$. Since $\alpha A \neq \alpha' A'$ or $x \neq w'$, it follows from derivations (A.1) and (A.2) that the production $A \to \varepsilon$ does not satisfy the $LR(k)$ condition. Thus G is not an $LC(k)$ grammar. This contradicts the assumption that G is such a grammar. We conclude that all ε-productions of G satisfy the A-$LR(k)$-condition.

Finally, suppose that G is not an A-$LC(k)$ grammar, because there is a production $A \to X\beta$, and (A.7) and (A.8) are derivations in G' such that $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$, although $\alpha'\alpha'' = \alpha$ and $k{:}y_1z_1 = k{:}y_2z_2$. Assume both $A, B \neq S'$. Then (A.5) and (A.6) are derivations in G, $\alpha'\alpha'' = \alpha$ and $k{:}y_1z_1 = k{:}y_2z_2$. Since $\alpha A \neq \alpha' B$ or $\beta \neq \gamma$, this implies that G is not an $LC(k)$ grammar. This contradicts the assumption that G is an $LC(k)$ grammar. Assume that $A = S'$. Then the production $A \to X\beta$ equals the production $S' \to S$ and derivation (A.7) has the form

$$S' \Rightarrow_r S. \tag{A.9}$$

Since in this case $\alpha = \alpha'\alpha'' = \varepsilon$, derivation (A.8) has the form

$$S' \Rightarrow_r^* B z_2 \Rightarrow_r S\gamma z_2 \Rightarrow_r^* S y_2 z_2. \tag{A.10}$$

Since y_1z_1 from derivation (A.7) equals ε in derivation (A.9), it follows from $k{:}y_1z_1 = k{:}y_2z_2$ that in derivation (A.10) $y_2z_2 = \varepsilon$ (notice that the condition $k \neq 0$ is essential here). Since the production $B \to S\gamma$ is not the production $S' \to S$, it follows from derivation (A.10) that $S \Rightarrow_r^* B \Rightarrow_r S\gamma \Rightarrow_r^* S$ is a derivation in G. This implies that G is ambiguous, contradicting the assumption that G is an $LC(k)$ grammar. In the same way, the assumption that $B = S'$ leads to a contradiction. We conclude that G is an $A\text{-}LC(k)$ grammar. □

Acknowledgment

The author wishes to thank Anton Nijholt (Vrije Universiteit, Brussels) for useful suggestions and both referees for pointing out some mistakes and obscure parts in the manuscript.

References

[1] A.V. Aho and J.D. Ullman, *The Theory of Parsing, Translation and Compiling, Vol. 1: Parsing* (Prentice-Hall, Englewood Cliffs, NJ, 1972).

[2] H.J.A. op den Akker, A left-corner property for context-free grammars, Memorandum INF-86-8, Dept. of Computer Science, Twente University, Enschede, The Netherlands, 1986.

[3] J.C. Beatty, Two iteration theorems for the LL(k) languages, *Theoret. Comput. Sci.* **12** (1980) 193–228.

[4] M.M. Geller and M.A. Harrison, On LR(k) grammars and languages, *Theoret. Comput. Sci.* **4** (1977) 245–276.

[5] M.A. Harrison, *Introduction to Formal Language Theory* (Addison-Wesley, Reading, MA, 1978).

[6] K.N. King, Iteration theorems for families of strict deterministic languages, *Theoret. Comput. Sci.* **10** (1980) 317–333.

[7] A.J. Korenjak and J.E. Hopcroft, Simple deterministic grammars, in: *Conf. Record 7th Ann. Symp. on Switching and Automata Theory* (1966) 36–46.

[8] D.J. Rosenkrantz and P.M. Lewis II, Deterministic left corner parsing, in: *IEEE Conf. Record 11th Ann. Symp. on Switching and Automata Theory* (1970) 139–152.

[9] E. Soisalon-Soininen, Characterization of LL(k) languages by restricted LR(k) grammars, Dept. of Computer Science, University of Helsinki, Helsinki, Report A-1977-3, 1977.

[10] E. Soisalon-Soininen and E. Ukkonen, A characterization of LL(k) languages, in: S. Michaelson and R. Milner, eds., *Automata, Languages and Programming* (Edinburgh University Press, Edinburgh, 1976) 20–30.

[11] E. Soisalon-Soininen and E. Ukkonen, A method for transforming grammars into LL(k) form, *Acta Inform.* **12** (1979) 339–369.