NOTES ON TOP-DOWN LANGUAGES

R. KURKI-SUONIO

Abstract.

Properties of context-free languages and grammars permitting deterministic top-down recognition with bounded lookahead are discussed. In particular, it is shown that for each \( k > 1 \) there are such languages requiring a lookahead of at least \( k \) characters.

Key words: Context-free, language, top-down, lookahead.

1. Introduction and summary.

The concept and definition of \( LL(k) \) grammars, which permit deterministic top-down recognition from left to right with a lookahead of \( k \) characters, have been introduced by Knuth in [1], and by Lewis and Stearns in [5]. The concept of an \( s \)-grammar, which is a special case of an \( LL(1) \) grammar is due to Korenjack and Hoperoft [2]. These concepts have been recently investigated also by Wood in [6].

Sections 2 and 3 of this paper contain definitions and simple theorems which are needed in the later sections.

In section 4 we give some necessary and sufficient properties of \( LL(k) \) languages and grammars. It is shown (theorem 7) that the family of \( LL(k) \) languages is generated by a proper subset of \( LL(k) \) grammars which allows recognition by one-state deterministic pushdown automata with a lookahead of \( k \) characters.

Section 5 deals with the question of reducing the amount of lookahead. It is found that reducing \( k \) by one or replacing an \( LL(1) \) grammar by an \( s \)-grammar might result in a grammar with shorter terminal phrases. This reduction is therefore not possible for some grammars containing null phrases (theorem 9). The special case for \( LL(1) \) and \( s \)-grammars has been treated earlier by the author in [4].

As some kind of a converse to this result it is shown (theorem 11) that if the null string is no sentence it can be deleted from the grammar with the effect of increasing \( k \) by one.

Received April 23, 1969; revised June 24, 1969.
2. Notations and definitions.

Single terminal and nonterminal symbols will be denoted by $a, b, \ldots,$ and $A, B, \ldots$. Strings of terminals, and strings containing terminals and/or nonterminals will be denoted by $x, y, \ldots,$ and $\alpha, \beta, \ldots$. By $\alpha^i$ ($i \geq 0$), we mean the string $\alpha \ldots \alpha$ in which $\alpha$ occurs $i$ times. If only part of a string is displayed, this is indicated by three dots, as e.g. in $x \ldots$, or $\ldots A x \ldots$. The designated nonterminal of a context-free grammar will be denoted by $S$, and the string of length zero by $\varepsilon$.

The collection of all productions $A \rightarrow \alpha_i$, $(i = 1, \ldots, n)$, for a nonterminal $A$ will also be written as $A \rightarrow \alpha_1 | \ldots | \alpha_n$, and it is called the production rule for $A$. Since we are in this paper interested only in such derivations in which productions are applied to the leftmost nonterminal of a string only, we define: $xAx \rightarrow \beta$ if and only if $\beta = xx'x$ such that $A \rightarrow x'$; and $\alpha \Rightarrow \beta$ if and only if either $\alpha = \beta$ or there is a derivation sequence $\alpha = \alpha_0 \rightarrow \alpha_1 \rightarrow \ldots \rightarrow \alpha_n = \beta$. The reader is warned that these notations usually have different meanings in the literature.

The arrows $\rightarrow$ and $\Rightarrow$ of this paper correspond to the notations $L$ and $L^*$ in [1].

The length of a string $\alpha$ will be denoted by $l(\alpha)$, and the length of the shortest terminal string derivable from $\alpha$ by $m(\alpha)$.

A nonterminal $A$ is called useless if the situation $S \Rightarrow \ldots A \ldots \Rightarrow x$ does not occur for any $x$. Without mentioning explicitly we shall assume in the following that a context-free grammar does not contain useless nonterminals.

Discussion on lookahead of $k$ characters is simplified if there are always at least $k$ characters to examine. For this reason we introduce a new terminal $\perp$ and a new non-terminal $S_k$ with one production $S_k \rightarrow S_{k+1}$. If the original grammar was called $G$, the modified grammar, with $S_k$ as the designated nonterminal, will be called $G_k$.

**Definition 1.** A production $A \rightarrow \alpha_1$ in a context-free grammar $G$ is called an $LL(k)$ production, if in $G_k$

$$S_k \Rightarrow xA\alpha \Rightarrow xx_1\alpha \Rightarrow xx \ldots,$$

$$S_k \Rightarrow xA\beta \Rightarrow xx_2\beta \Rightarrow xx \ldots, \quad l(z) = k,$$

implies $\alpha_1 = \alpha_2$.

**Definition 2.** A nonterminal (context-free grammar) is called an $LL(k)$ nonterminal (grammar), if all its productions are $LL(k)$ productions.