INFORMATION AND CONTROL 53, 52–65 (1982)

Derivational Complexity of Context-Free Grammars

SEPPO SIPPU*

University of Helsinki, Helsinki, Finland

Derivational complexity of context-free grammars is studied. Minimal grammar-dependent upper bounds are determined both on the derivational time complexity, that is, the number of derivation steps needed to derive a sentence of given length, and on the derivational space complexity, that is, the length of the longest sentential form needed in the derivation. In addition to general context-free grammars, these upper bounds are also determined specifically for ε-free grammars, non-left-recursive and non-right-recursive grammars, and for LL(k) grammars. The results might prove useful in parser optimization, because the complexity of a parser is closely related to the derivational complexity of the underlying context-free grammar.

1. INTRODUCTION

The context-free grammars of programming languages are usually required to be "parsable," that is, LL(k) or LR(k) grammars or their variants. These grammars have the virtue of possessing deterministic parsing algorithms of a linear time complexity. That is, for each sentence w derived by the grammar a parse tree can be produced deterministically in time $c|w| + d$, where $|w|$ is the length of w, and c and d are constants that depend only on the grammar in question.

There is a close correspondence between the time complexity of a parser and the *derivational* time complexity of the underlying context-free grammar, that is, the number of derivation steps needed to derive a sentence in the language. In processing a sentence w, an LR(k) parser, for example, performs as many reduce actions by grammar productions as is needed to construct the parse tree for w. Similarly, an LL(k) parser performs the same number of produce actions. In addition, both parsers perform $|w|$ shift actions. Thus, the time complexity of a parser normally equals $|w|$ plus the derivational time complexity of the underlying grammar. An almost equally

* The work was supported by the Academy of Finland. Present address: Department of Mathematics and Physics, University of Joensuu, P.O. Box 111, SF-80101 Joensuu 10, Finland.

0019-9958/82 $2.00

close correspondence exists between the space complexity of a parser and the derivational space complexity of the grammar, that is, the length of the longest sentential form needed in the derivation.

In the literature on parsing theory, a great deal of effort has been made to develop optimization methods that decrease the complexity of a parser by some constant factor (see, e.g., Aho and Ullman, 1973 and Pager, 1977). As the problem of parser optimization reduces in a simple way to the problem of decreasing the complexity of the underlying grammar, this invites us to determine the exact complexity of context-free grammars, that is, how the constants c and d in the complexity bound $c|w| + d$ depend on the grammar in question and how they vary when different classes of parsable grammars are considered.

In the present paper we determine tight time and space bounds on the grammar-dependent constants c and d in the cases in which the grammar in question belongs to one of the following classes: (1) general context-free grammars, (2) ε-free grammars, (3) non-left-recursive grammars, (4) non-right-recursive grammars, and (5) LL(k) grammars. The bounds given are *minimal* in that in each case there is a sequence of grammars for which the bounds are actually reached in the derivation of every sentence.

We begin the presentation with ε-free grammars, and determine their complexity in Section 2. In Section 3 we determine the complexity of deriving the empty string ε, and in Section 4 we combine these results to obtain the complexity of general context-free grammars. In Section 5 we consider the complexity of leftmost and rightmost derivations, and in Section 6 we make some concluding remarks and pose some open problems.

In the literature, the notion of derivational complexity has gained little attention. Even the fundamental fact that the time complexity of context-free grammars (and parsers of parsable grammars) indeed is linear in the length of the sentence is usually taken for granted and mentioned without proof. Aho and Ullman (1972) deduced the linear time complexity of canonical LR(k) parsers indirectly from a result concerning looping configurations in general push-down automata. Hopcroft and Ullman (1979) mentioned the linear time complexity of context-free grammars as a "starred" exercise. Harrison (1978) and Heilbrunner (1981) proved the linear time complexity for context-free grammars, but their upper bounds were not the best possible.

Book (1971) considered derivational complexity from a language theoretic point of view and developed a general complexity theory by using grammars in place of Turing machines (also see Salomaa, 1973). From the point of view of pure language theory, the notion of derivational complexity is rather trivial in the case of context-free grammars, because any (nonempty) sentence w in a context-free language can always be derived in *real* time, that is, in $|w|$ derivation steps, if a Greibach normal-form grammar is used. However, this result is not so useful in parsing and compiling theory, where

we are interested not only in the language generated but also in the particular grammar used.

We conclude this section by reviewing some basic concepts concerning strings and context-free grammars. We make free use of the notations and definitions given in Aho and Ullman (1972). We stipulate that A, B, and C denote nonterminals, a and b denote terminals, X denotes either a nonterminal or a terminal, w denotes a terminal string, α and ω denote general strings, π denotes a production string, and ε denotes the empty string. The "(general) derives" relation of a grammar G is denoted by $\Rightarrow$. If π is a production string, $\Rightarrow^{\pi}$ denotes the derives relation that uses the production string π. The language generated by a symbol X of G is denoted by $L_G(X)$, or $L(X)$, for short. A grammar is *ε-free* if it has no production with an empty right-hand side. A grammar G is *left-recursive* (resp. *right-recursive*) if $A \Rightarrow^{+} A\alpha$ (resp. $A \Rightarrow^{+} \alpha A$) holds in G for some nonterminal A and string α.

A sequence of strings $\langle \alpha_0, \ldots, \alpha_n \rangle$ is a *derivation of α_n from α_0 in G* if $\alpha_i \Rightarrow \alpha_{i+1}$ holds in G for all $i = 0, \ldots, n-1$. We say that n is the *time complexity*, and $\max\{|\alpha_i| \mid i = 0, \ldots, n\}$ the *space complexity*, of the derivation. String α *derives* string α' *in time t* (resp. *in space s*) if α' has a derivation from α of time complexity at most t (resp. space complexity at most s). α *derives α' simultaneously in time t and in space s* if α' has a derivation from α of time complexity at most t and space complexity at most s. By the *time complexity* (resp. *space complexity*) *of deriving α' from α* we mean the least integer n such that α derives α' in time n (resp. in space n). For convenience, we take the liberty of speaking of "derivations" $\alpha \overset{*}{\Rightarrow} \alpha'$ and of their time and space complexities when we actually mean the corresponding string sequences $\langle \alpha, \ldots, \alpha' \rangle$.

2. Complexity of ε-Free Grammars

Let $m \geqslant 1$ and let G_m be the grammar with the productions

$$\begin{aligned} A_1 &\to A_2 \\ A_2 &\to A_3 \\ &\vdots \\ A_m &\to a \mid A_1 A_1 . \end{aligned}$$

$L(A_1) = a^{+}$ and A_1 derives the sentence a^k (i.e., the string of k a's) simultaneously in time $2mk - m$ and in space k, for all $k \geqslant 1$.

In fact, we have

THEOREM 1. *Let G be an ε-free grammar with m nonterminals. If X is a symbol of G and w is in $L(X)$, then X derives w simultaneously in time*

$$2m|w| - m$$

and in space

$$|w|.$$

Moreover, these bounds are minimal.

Proof. First we note that the space complexity of any derivation in G is $|w|$, because in an ε-free grammar no derivation step can decrease the length of the sentential form.

We now prove, by induction on $|w|$, that whenever w is in $L(X)$, then X derives w in time $2m|w| - m$. If $|w| = 1$ and $X \Rightarrow^{\pi} w$, then the ε-freedom of G implies that π can contain only unit productions of the form $A \to B$ or $B \to w$. Thus, if π is the shortest possible production string such that $X \Rightarrow^{\pi} w$, then $|\pi| \leqslant m = 2m|w| - m$, because the appearance in π of two productions of the same nonterminal would imply an unnecessary loop.

We may thus assume that $|w| > 1$ and, as an induction hypothesis, that whenever w' is in $L(X')$ and $|w'| < |w|$, then X' derives w' in time $2m|w'| - m$. If $X \overset{*}{\Rightarrow} w$, then there is a production $r = A \to X_1 \cdots X_n$, $n \geqslant 2$, production strings $\pi, \pi_1, \ldots, \pi_n$ and terminal strings $w_1, \ldots, w_n$ such that

$$X \overset{\pi}{\Longrightarrow} A \overset{r}{\Longrightarrow} X_1 \cdots X_n, \qquad X_i \overset{\pi_i}{\Longrightarrow} w_i \quad \text{for all } i, \quad \text{and } w_1 \cdots w_n = w.$$

Here π can contain only unit productions of the form $B \to C$. Thus, if π is the shortest possible, then $|\pi| \leqslant m - 1$. On the other hand, $n \geqslant 2$ implies that $|w_i| < |w|$ for all $i = 1, \ldots, n$, which means that we can apply the induction hypothesis and assume that $|\pi_i| \leqslant 2m|w_i| - m$ for all $i = 1, \ldots, n$.

We then have

$$\begin{aligned} |\pi r \pi_1 \cdots \pi_n| &= |\pi| + 1 + \sum_{i=1}^{n} |\pi_i| \\ &\leqslant (m-1) + 1 + \sum_{i=1}^{n} (2m|w_i| - m) \\ &= 2m|w| + (1-n)m \\ &\leqslant 2m|w| - m, \end{aligned}$$

as claimed. Moreover, the grammars G_m, $m \geqslant 1$, presented above show that this bound is also minimal. ∎

The grammars G_m, $m \geqslant 1$, are both left and right recursive. It turns out

that this is a necessary property of the grammar, if the bound $2m|w| - m$ is to be actually reached. Before showing this, we consider the grammar G'_m with the productions

$$\begin{aligned} A_1 &\to A_2 \\ A_2 &\to A_3 \\ &\vdots \\ A_m &\to aA_1 \mid b. \end{aligned}$$

G'_m is an LL(1) grammar, $L(A_1) = a^*b$, and A_1 derives the sentence a^kb in time $m(k+1)$, for all $k \geqslant 0$.

THEOREM 2. *Let G be a non-left-recursive or non-right-recursive ε-free grammar with m nonterminals. If X is a symbol of G and w is in $L(X)$, then X derives w in time*

$$m|w|.$$

Moreover, this bound is minimal.

Proof. First we note that the non-right-recursive case follows immediately from the non-left-recursive case, because G is right-recursive if and only if its *reversed* grammar G^R is left-recursive. Here G^R is obtained from G by replacing each production $A \to \omega$ in G by $A \to \omega^R$ in which ω^R is the reversal, or mirror image, of ω. Clearly, X derives w in G in time t if and only if X derives w^R in G^R in time t.

We now prove, by induction on $|w|$, the non-left-recursive case. The case $|w| = 1$ is proved as in Theorem 1. In the case $|w| > 1$ the condition $X \overset{*}{\Rightarrow} w$ and the non-left-recursiveness and the ε-freedom of G imply that there is a terminal a, symbols $X_1, \ldots, X_n$, production strings $\pi, \pi_1, \ldots, \pi_n$ and terminal strings $w_1, \ldots, w_n$ such that

$$X \overset{\pi}{\Longrightarrow} aX_1 \cdots X_n, \quad |\pi| \leqslant m, \qquad X_i \overset{\pi_i}{\Longrightarrow} w_i \quad \text{for all } i, \qquad \text{and } aw_1 \cdots w_n = w.$$

Thus, if $|\pi_i| \leqslant m|w_i|$ for all i, we have

$$\begin{aligned} |\pi\pi_1 \cdots \pi_n| &= |\pi| + \sum_{i=1}^{n} |\pi_i| \\ &\leqslant m + \sum_{i=1}^{n} m|w_i| \\ &= m|w|, \end{aligned}$$

as claimed. The grammars G'_m, $m \geqslant 1$ (and their reversed grammars) show that this bound is also minimal. ∎

3. Complexity of Deriving the Empty String

Let $m' \geqslant 1$, $n \geqslant 2$ and let $G_{m',n}$ be the grammar with the productions (A_i^n means the string of n A_i's)

$$A_1 \to A_2^n$$
$$A_2 \to A_3^n$$
$$\vdots$$
$$A_{m'} \to \varepsilon.$$

$G_{m',n}$ is an LL(0) grammar and $L(A_i) = \{\varepsilon\}$ for all $i = 1,..., m'$. Moreover, A_i derives ε simultaneously in time

$$1 + n + \cdots + n^{k_i - 1} = \frac{n^{k_i} - 1}{n - 1}$$

and in space

$$(k_i - 1)(n - 1) + 1,$$

where $k_i = m' - i + 1$. Note that k_i is the height of the parse tree that corresponds to the derivation $A_i \overset{*}{\Rightarrow} \varepsilon$.

To prove that these time and space complexities in fact are upper bounds on the time and space complexities of deriving ε in any grammar, we define the following sets:

$$V = \{A \mid A \text{ is a } \textit{nullable} \text{ nonterminal, i.e., } A \Rightarrow^+ \varepsilon\},$$
$$V_1 = \{A \mid \text{the grammar has the production } A \to \varepsilon\},$$

and, for all $k > 1$,

$$V_k = \Big\{A \mid A \notin \bigcup_{i=1}^{k-1} V_i, \text{ and the grammar has a production } A \to A_1 \cdots A_l$$
$$\text{in which each } A_i \in V_{k_i} \text{ for some } k_i < k\Big\}.$$

Intuitively, A is in V_k if and only if A derives ε by a parse tree of height k, but not by any parse tree of height $k' < k$.

We have

Lemma 3. *$V = \bigcup_{k=1}^{m'} V_k$, where $m' = |V|$, i.e., the number of nullable nonterminals in the grammar.*

Proof. First, it is clear that each V_k is included in V. On the other hand, if $A \Rightarrow^j \varepsilon$, we can show by a simple induction on j that A is in V_k for some k. Thus, $V = \bigcup_{k=1}^{\infty} V_k$.

If $k > 1$ and A is in V_k, then, by definition, $A \notin \bigcup_{i=1}^{k-1} V_i$ and the grammar has a production $A \to A_1 \cdots A_l$ in which each A_i is in V_{k_i} for some $k_i < k$. Here $k_i = k-1$ for some i, because otherwise the conditions $k_i < k-1$, $i = 1,..., l$, would imply that A is in $V_{k'}$, for some $k' \leqslant k-1$. Thus, for all $k > 1$ V_{k-1} is nonempty whenever V_k is nonempty. In other words, for all $k \geqslant 1$ V_{k+1} is empty whenever V_k is empty. Because the sets V_k form a pairwise disjoint partition of the set V, we can conclude that V_k is empty for all $k > m'$. ∎

LEMMA 4. *Let G be a grammar and $n \geqslant 2$ the length of the right-hand side of the longest production in G. Then for all $k \geqslant 1$ and A in V_k, A derives ε simultaneously in time*

$$\frac{n^k - 1}{n - 1}$$

and in space

$$(k-1)(n-1) + 1.$$

Moreover, these bounds are minimal.

Proof. The proof is by induction on k. If $k = 1$, $A \to \varepsilon$ is a production of G. Thus, A derives ε simultaneously in time $1 = (n^k - 1)/(n-1)$ and in space $1 = (k-1)(n-1) + 1$. We can therefore assume that $k > 1$ and, as an induction hypothesis, that whenever $k' < k$ and A' is in $V_{k'}$, then A' derives ε simultaneously in time $(n^{k'} - 1)/(n-1)$ and in space $(k'-1)(n-1) + 1$. By definition, G has a production $r = A \to A_1 \cdots A_l$ in which each A_i is in V_{k_i} for some $k_i < k$. By the induction hypothesis, there are production strings $\pi_1,..., \pi_l$ such that, for all $i = 1,..., l$, $A_i \Rightarrow^{\pi_i} \varepsilon$, $|\pi_i| \leqslant (n^{k_i} - 1)/(n-1)$ and the space complexity of each of these derivations, denoted by s_i, is at most $(k_i - 1)(n-1) + 1$. We have

$$A \overset{r}{\Longrightarrow} A_1 \cdots A_l \overset{\pi_1}{\Longrightarrow} A_2 \cdots A_l \overset{\pi_2}{\Longrightarrow} \cdots \overset{\pi_{l-1}}{\Longrightarrow} A_l \overset{\pi_l}{\Longrightarrow} \varepsilon.$$

The time complexity of this derivation is

$$1 + \sum_{i=1}^{l} |\pi_i| \leqslant 1 + \sum_{i=1}^{l} \frac{n^{k_i} - 1}{n-1} \leqslant 1 + l\frac{n^{k-1} - 1}{n-1}$$

$$\leqslant 1 + n\frac{n^{k-1} - 1}{n-1} = \frac{n^k - 1}{n-1},$$

as claimed. The space complexity of the derivation is

$$\begin{aligned}
&\max\{s_i + (l-i) \mid i = 1,\ldots, l\} \\
&\quad \leqslant \max\{(k_i - 1)(n-1) + 1 + (l-i) \mid i = 1,\ldots, l\} \\
&\quad \leqslant \max\{((k-1)-1)(n-1) + 1 + (l-i) \mid i = 1,\ldots, l\} \\
&\quad \leqslant (k-2)(n-1) + 1 + (l-1) \\
&\quad \leqslant (k-2)(n-1) + 1 + (n-1) \\
&\quad = (k-1)(n-1) + 1,
\end{aligned}$$

as claimed. The grammars $G_{m',n}$, $m' \geqslant 1$, $n \geqslant 2$, given above show that these bounds are also minimal. ∎

By Lemmas 3 and 4 we have

THEOREM 5. *Let G be a grammar, $m' \geqslant 1$ the number of nullable nonterminals in G, and let $n \geqslant 2$ be the length of the right-hand side of the longest production in G. Then any nullable nonterminal A in G derives ε simultaneously in time*

$$\frac{n^{m'} - 1}{n - 1}$$

and in space

$$(m' - 1)(n - 1) + 1.$$

Moreover, these bounds are minimal.

4. Complexity of General Context-Free Grammars

Let $m \geqslant 1$, $m' \geqslant 1$, $n \geqslant 2$, and let $G_{m,m',n}$ be the grammar with the productions

$$\begin{array}{ll}
A_1 \to A_2 B_1^{n-1} & B_1 \to B_2^n \\
A_2 \to A_3 B_1^{n-1} & B_2 \to B_3^n \\
\quad\vdots & \quad\vdots \\
A_m \to A_1 A_1 B_1^{n-2} \mid a B_1^{n-1} & B_{m'} \to \varepsilon.
\end{array}$$

$G_{m,m',n}$ has been obtained by combining the grammars G_m and $G_{m',n}$ (see Sections 2 and 3).

If $ck + d$ is the time complexity of deriving a^k from A_1 in G_m and if t is the time complexity of deriving ε from B_1, then the time complexity of deriving a^k from A_1 in $G_{m,m',n}$ is

$$(ck+d)+((ck+d)(n-1)-(k-1))t$$
$$=(c+c(n-1)t-t)k+d+d(n-1)t+t.$$

Note that any application of an A-production introduces in the sentential form $n-1$ instances of B_1, except for the production $A_m \to A_1 A_1 B_1^{n-2}$, which introduces only $n-2$ instances. In the derivation of the sentence a^k this production is used $k-1$ times.

The most space-efficient way to derive a^k is to erase the B_1's from the sentential form as soon as they appear. That is, after each application of an A-production the introduced $n-1$ or $n-2$ instances of B_1 are let to derive ε in the most space-efficient way. Thus, if s is the space complexity of deriving ε from B_1, then the space complexity of deriving a^k from A_1 in $G_{m,m',n}$ is

$$k+(n-2)+s.$$

To show that these time and space complexities are upper bounds on the time and space complexities of any context-free grammar, we use the well-known method for removing ε-productions from general context-free grammars. If G is a grammar, we denote by $\hat{G}$ the ε-free grammar that has been obtained from G by replacing the production set P of G by the set

$$\hat{P}=\{A \to \alpha_1\alpha_2 \cdots \alpha_l\alpha_{l+1} \mid l \geqslant 0, \alpha_1\alpha_2 \cdots \alpha_l\alpha_{l+1} \neq \varepsilon \text{ and for some nullable nonterminals } B_1, B_2, \ldots, B_l \text{ of } G, A \to \alpha_1 B_1 \alpha_2 B_2 \cdots \alpha_l B_l \alpha_{l+1} \text{ is in } P\}.$$

It can be shown (see, e.g., Aho and Ullman, 1972 or Hopcroft and Ullman, 1979) that $L_{\hat{G}}(A)=L_G(A)\backslash\{\varepsilon\}$ for all nonterminals A. Note that in the case of the grammar $G_{m,m',n}$ the transformation produces a grammar $\hat{G}_{m,m',n}$ in which the set of *useful* productions (i.e., those productions that can be used in the derivation of some sentence) is exactly the production set of G_m.

THEOREM 6. *Let G be a grammar in which the length of the right-hand side of the longest production is n, $n \geqslant 2$, and in which each nullable nonterminal derives ε simultaneously in time t and in space s. If in the corresponding ε-free grammar $\hat{G}$ a nonterminal A derives a terminal string w in time $c|w|+d$, then in G A derives w simultaneously in time*

$$(c+c(n-1)t-t)|w|+d+d(n-1)t+t$$

and in space

$$|w|+(n-2)+s.$$

Moreover, these bounds are minimal.

Proof. For each nullable nonterminal B in G, let $D(B)$ be a production string such that B derives ε by using $D(B)$ simultaneously in time t and in space s. Furthermore, for each production $A \to \omega$ in $\hat{G}$ we choose a production $A \to \alpha_1 B_1 \cdots \alpha_l B_l \alpha_{l+1}$ in G such that $\alpha_1 \cdots \alpha_{l+1} = \omega$ and each B_i is a nullable nonterminal. The construction of $\hat{G}$ guarantees that this choice is possible. We then define homomorphisms f and g from $\hat{P}^*$ to P^* as follows:

$$f(A \to \omega) = A \to \alpha_1 B_1 \cdots \alpha_l B_l \alpha_{l+1},$$
$$g(A \to \omega) = f(A \to \omega)\, D(B_1) \cdots D(B_l).$$

We say that $A \to \omega$ has l *ε-positions* and $|\omega|$ *non-ε-positions* (with respect to f).

Now if $A \Rightarrow^{\pi} w$ in $\hat{G}$, then it can be shown by a simple induction on $|\pi|$ that $A \Rightarrow^{g(\pi)} w$ in G (actually, the proof of this is part of the proof of the fact that $L_{\hat{G}}(A) = L_G(A) \backslash \{\varepsilon\}$). Note that this derivation is obviously the least space consuming because the nullable B's are erased as soon as they appear.

The time complexity of the derivation $A \Rightarrow^{g(\pi)} w$ is at most $|\pi| + et$ in which e is the total number of ε-positions in the productions in π. Because $\hat{G}$ is ε-free and because the length of the right-hand side of any production in G is at most n, each production in π has at least one non-ε-position and at most $n-1$ ε-positions. Thus at least $e \geqslant |\pi|(n-1)$. However, if $|w| > 1$, not all of the productions in π can be unit productions. More precisely, there must be $|w| - 1$ additional non-ε-positions in π. This means that actually $e \leqslant |\pi|(n-1) - (|w| - 1)$. If $|\pi| \leqslant c\,|w| + d$, we can thus conclude that the time complexity of the derivation $A \Rightarrow^{g(\pi)} w$ is at most

$$c\,|w| + d + ((c\,|w| + d)(n-1) - (|w| - 1))t$$
$$= (c + c(n-1)t - t)\,|w| + d + d(n-1)t + t,$$

as claimed.

Since after any application of a production $f(A \to \omega)$ the nullable nonterminals that correspond to the ε-positions in $A \to \omega$ are immediately erased, one by one, in space s, no immediate sentential form in the derivation $A \Rightarrow^{g(\pi)} w$ can contain more than $(n-2) + s$ nullable nonterminals that arise from ε-positions. This means that the space complexity of the derivation is at most $|w| + (n-2) + s$, as claimed.

The minimality of the bounds follows from the grammars $G_{m,m',n}$. ∎

By combining Theorems 1, 5, and 6 we get

THEOREM 7. *Let G be a grammar, $m \geqslant 1$ the number of nonterminals that derive a nonempty terminal string, $m' \geqslant 0$ the number of nullable*

nonterminals, and let $n \geqslant 2$ be the length of the right-hand side of the longest production in G. If A is a nonterminal and w is in $L(A)\backslash\{\varepsilon\}$, then A derives w simultaneously in time

$$\left(2mn^{m'} - \frac{n^{m'}-1}{n-1}\right)|w| - mn^{m'} + \frac{n^{m'}-1}{n-1}$$

and in space

$$|w| + m'(n-1).$$

Moreover, these bounds are minimal.

By combining Theorems 2, 5, and 6 we get

THEOREM 8. *Let G be a non-left-recursive or non-right-recursive grammar, $m \geqslant 1$ the number of nonterminals that derive a nonempty terminal string, $m' \geqslant 0$ the number of nullable nonterminals, and let $n \geqslant 2$ be the length of the right-hand side of the longest production in G. If A is a nonterminal and w is in $L(A)\backslash\{\varepsilon\}$, then A derives w simultaneously in time*

$$\left(mn^{m'} - \frac{n^{m'}-1}{n-1}\right)|w| + \frac{n^{m'}-1}{n-1}$$

and in space

$$|w| + m'(n-1).$$

Moreover, these bounds are minimal.

The minimality of the bounds given in Theorem 8 can be seen by combining the grammars G'_m (see Section 2) and $G_{m',n}$ to yield the following grammar, $G'_{m,m',n}$:

$$\begin{array}{ll}
A_1 \to A_2 B_1^{n-1} & B_1 \to B_2^n \\
A_2 \to A_3 B_1^{n-1} & B_2 \to B_3^n \\
\quad\vdots & \quad\vdots \\
A_m \to aA_1 B_1^{n-2} \mid bB_1^{n-1} & B_{m'} \to \varepsilon.
\end{array}$$

Because no LL(k) grammar can be left-recursive and because the grammar $G'_{m,m',n}$ is an LL(1) grammar for all m, m', and n, we have

COROLLARY 9. *The bounds given in Theorem 8 are minimal upper bounds on the derivational time and space complexity of* LL(k) *grammars, $k \geqslant 1$.*

5. Complexity of Leftmost and Rightmost Derivations

So far, we have considered only general derivational complexity of grammars. That is, given A and w such that $A \overset{*}{\Rightarrow} w$ we have determined tight time and space bounds on the most efficient derivation of w from A. In parsing theory we are, however, usually concerned with special kinds of derivations, such as "leftmost" or "rightmost" derivations. An LL(k) parser always produces a leftmost derivation for the sentence to be parsed, which means that its complexity is directly related to the complexity of leftmost deriving, rather that general deriving, in the underlying grammar. The complexity of an LR(k) parser is similarly related to the complexity of rightmost deriving.

Theorems 1, 2, and 5 are easily seen to hold for leftmost and rightmost derivations as well. So are the time bounds given in Theorems 6–8. Note that if $A \Rightarrow^{\pi} w$, then $A \Rightarrow_{lm}^{\pi'} w$ and $A \Rightarrow_{rm}^{\pi''} w$ for some permutations π' and π'' of π. Here $\Rightarrow_{lm}^{\pi'}$ and $\Rightarrow_{rm}^{\pi''}$ denote the "leftmost derives" and "rightmost derives" relations that use the production strings π' and π'', respectively. However, the space bounds in Theorems 6–8 have been obtained by using derivations that obviously are neither leftmost nor rightmost. Indeed, in any leftmost derivation of the sentence $a^k b$ in the grammar $G'_{m,m',n}$ (see the previous section) the erasing of the nullable nonterminals B_1 can begin only at the left sentential form $a^k b B_1^e$ in which e is the total number of B_1's in the A-productions that have been applied to produce $a^k b B_1^e$. Clearly, $e = (c(k+1)+d)(n-1)-k$ in which $c(k+1)+d$ is the time complexity of deriving $a^k b$ in G'_m (see Section 2). Thus, the space complexity of leftmost deriving $a^k b$ in $G'_{m,m',n}$ is

$$(k+1)+(e-1)+s = c(n-1)(k+1)+d(n-1)+s,$$

in which s is the space complexity of deriving ε from B_1.

Theorem 10. *Let G be a grammar in which the length of the right-hand side of the longest production is n, $n \geqslant 2$, and in which each nullable nonterminal leftmost (resp. rightmost) derives ε in space s. If in the corresponding ε-free grammar $\hat{G}$ a nonterminal A leftmost (resp. rightmost) derives a terminal string w in time $c|w|+d$, then in G A leftmost (resp. rightmost) derives w in space*

$$c(n-1)|w|+d(n-1)+s.$$

Proof. We consider only leftmost derivations; the rightmost derivations are handled analogously. For each nullable nonterminal B in G, let $D(B)$ be a production string such that B leftmost derives ε by using $D(B)$ in space s.

Furthermore, let f be the homomorphism defined in the proof of Theorem 6. If $A \Rightarrow_{lm}^{\pi} w$ in $\hat{G}$, then, by the definition of f, we have in G

$$\begin{aligned} A &\xRightarrow{f(\pi)} w_1 B_1 w_2 B_2 \cdots w_e B_e w_{e+1} \\ &\xRightarrow[lm]{D(B_1)} w_1 w_2 B_2 \cdots w_e B_e w_{e+1} \\ &\quad\vdots \\ &\xRightarrow[lm]{D(B_e)} w_1 w_2 \cdots w_e w_{e+1} = w. \end{aligned}$$

Here e is the total number of ε-positions in π, and $B_1, \ldots, B_e$ are the nullable nonterminals that correspond to these ε-positions. The first derivation segment, i.e., that using $f(\pi)$, is a leftmost derivation except that the nullable B's are not touched. Thus, the derivation is at least as space consuming as the most space-efficient leftmost derivation $A \overset{*}{\Rightarrow}_{lm} w$ in G. The space complexity of the derivation is $|w| + (e-1) + s$. Thus, if $|\pi| \leqslant c|w| + d$, we can conclude that the space complexity of leftmost deriving w from A in G is at most

$$\begin{aligned} &|w| + (e-1) + s \\ &\quad \leqslant |w| + ((c|w| + d)(n-1) - (|w| - 1)) - 1) + s \\ &\quad = c(n-1)|w| + d(n-1) + s, \end{aligned}$$

as claimed. Note that $e \leqslant |\pi|(n-1) - (|w| - 1)$ as in the proof of Theorem 6. ∎

By combining Theorems 2, 5, and 10 we get

THEOREM 11. *Let G be a non-left-recursive (resp. non-right-recursive) grammar, $m \geqslant 1$ the number of nonterminals that derive a nonempty terminal string, $m' \geqslant 0$ the number of nullable nonterminals, and let $n \geqslant 2$ be the length of the right-hand side of the longest production in G. If A is a nonterminal and w is in $L(A)\backslash\{\varepsilon\}$, then A leftmost (resp. rightmost) derives w in space*

$$m(n-1)|w| + (m'-1)(n-1) + 1.$$

Moreover, this bound is minimal.

COROLLARY 12. *The bound given in Theorem 11 is a minimal upper bound on the space complexity of leftmost deriving in* LL(k) *grammars, $k \geqslant 1$.*

6. Conclusions

We have studied derivational complexity from a parsing theoretic point of view, and determined minimal grammar-dependent upper bounds on the time and space complexity of deriving a sentence in a context-free grammar. The classes of grammars considered were the general context-free grammars, the ε-free grammars, the non-left-recursive grammars, the non-right-recursive grammars and the LL(k) grammars. Moreover, we determined a minimal upper bound on the space complexity of leftmost deriving in non-left-recursive grammars and in LL(k) grammars, and on the space complexity of rightmost deriving in non-right-recursive grammars.

Several open problems remain. First, the time bound obtained for general context-free grammars seems to be reachable only in the case of ambiguous grammars. This suggests that a tighter bound might be possible to establish for unambiguous grammars. Second, the only class of parsable grammars considered was the LL(k) grammars. The bounds obtained for non-left-recursive grammars turned out to be minimal also for LL(k) grammars, $k \geqslant 1$. Unfortunately, we see no immediate way to obtain minimal bounds for the most important class of parsable grammars, the LR(k) grammars. Third, we leave open the problem of determining minimal space bounds for leftmost and rightmost deriving in general context-free grammars.

Acknowledgment

I wish to thank Eljas Soisalon-Soininen for illuminating discussions on the topic of the paper.

References

Aho, A. V., and Ullman, J. D. (1972), "The Theory of Parsing, Translation and Compiling. Vol. 1. Parsing," Prentice–Hall, Englewood Cliffs, N.J.

Aho, A. V., and Ullman, J. D. (1973), A technique for speeding up LR(k) parsers, *SIAM J. Comput.* **2**, 106–127.

Book, R. V. (1971), Time-bounded grammars and their languages, *J. Comput. System Sci.* **5**, 397–429.

Harrison, M. A. (1978), "Introduction to Formal Language Theory," Addison–Wesley, Reading, Mass.

Heilbrunner, S. (1981), A parsing automata approach to LR theory, *Theor. Comput. Sci.* **15**, 117–157.

Hopcroft, J. E., and Ullman, J. D. (1979), "Introduction to Automata Theory, Languages and Computation," Addison–Wesley, Reading, Mass.

Pager, D. (1977), Eliminating unit productions from LR parsers, *Acta Inform.* **9**, 31–59.

Salomaa, A. (1973), "Formal Languages," Academic Press, New York.