A Direct Proof of the Inherent Ambiguity of a Simple Context-Free Language

HERMAN A. MAURER

University of Calgary, Calgary, Alberta, Canada*

ABSTRACT. A direct and self-contained proof is given of the inherent ambiguity of the context-free language $L = \{a^ib^ic^j \mid i,j \geqslant 1\} \cup \{a^ib^jc^j \mid i,j \geqslant 1\}$, which is the solution to an open problem pointed out by Ginsburg.

KEY WORDS AND PHRASES: ambiguity, inherent ambiguity, unambiguity, grammars, context-free languages, Chomsky languages, phrase structure languages, production systems, type 2 languages, bounded languages

CR CATEGORIES: 5.22, 5.24

The existence of context-free, inherently ambiguous languages was first proved by Parikh [6] in 1961 by giving a direct proof of the inherent ambiguity of

$$M = \{a^ib^ja^ib^k \mid i, j, k \geq 1\} \cup \{a^ib^ja^kb^j \mid i, j, k \geq 1\}.$$

Later results concerning the ambiguity of context-free languages have mainly been obtained using powerful theorems on linear sets and bounded languages (see, e.g [3–5]). In his book [3] Ginsburg points out the open problem of finding a direct proof of the inherent ambiguity of $L = \{a^ib^ic^j \mid i,j \geq 1\} \cup \{a^ib^jc^j \mid i,j \geq 1\}$ [3, p. 211]. In this note a rather straightforward and direct proof is given of the inherent ambiguity of this context-free language L.

Preliminaries. A *context-free* (CF) *grammar*[1] G is a quadruple $G = (V_N, V_T, R, A)$ where V_N and V_T are finite sets (called *nonterminal alphabet* and *terminal alphabet*, respectively); A is an element of V_N; and R is a finite set of rules of the form $B \to \psi$ where B is an element of V_N, and ψ is a nonempty word[2] over $V = V_N \cup V_T$. A word ψ is said to be *immediately derived* from ξ, i.e. $\xi \underset{G}{\to} \psi$ or $\xi \longrightarrow \psi$, if $\xi = \rho B\delta$, $\psi = \rho\omega\delta$, and $B \to \omega$ is a rule of R. A word ψ is said to be *derived* from ξ (with respect to G); i.e. $\xi \underset{G}{\Rightarrow} \psi$ or $\xi \Rightarrow \psi$, if there exist $n \geq 1$ words $\xi_0, \xi_1, \cdots, \xi_n$ such that $\xi = \xi_0$, $\xi_n = \psi$, and $\xi_{i-1} \to \xi_i$ for $i = 1, 2, \cdots, n$. The sequence $\xi_0, \xi_1, \cdots, \xi_n$ is called a *ξ-derivation* of ψ (with respect to G) of *length* n. It is called a *leftmost* ξ-derivation of ψ (with respect to G) if $\xi_i = x_iB_i\delta_i$, $\xi_{i+1} = x_i\omega_i\delta_i$, and $B_i \underset{G}{\to} \omega_i$ for $i = 0, 1, 2, \cdots, n - 1$. It is well known [3] that for each

* Department of Mathematics. The research for this paper was supported by Grant A4008 of the National Research Council of Canada.

[1] It is a type 2 grammar according to [2].

[2] In this paper we adopt the same conventions used in [1–2]: capital letters denote words over V_N, lowercase letters denote words over V_T, and Greek letters denote words over V. Early letters of the alphabet denote individual symbols; late letters denote arbitrary (possibly empty) words. The symbol ϵ is reserved for denoting the empty word.

ξ-derivation of x with length n there exists a leftmost ξ-derivation of x of the same length. The set of all terminal words x for which there is an A-derivation is called the *CF language generated* by G and is denoted by $L(G)$. A CF grammar is called *unambiguous* if for each word $x \in L(G)$ there exists exactly one leftmost A-derivation with respect to G. Otherwise, G is called *ambiguous*. A CF language L is called *inherently ambiguous* if every CF grammar generating L is ambiguous. A CF grammar $G = (V_N, V_T, R, A)$ is called *reduced* if for each $B \in V_N$, $B \Rightarrow x$ for some x and for each $B \in V_N - \{A\}$, $A \Rightarrow yBz$ for some y and z. It is well known [3] that for each unambiguous CF grammar G there exists a reduced, unambiguous CF grammar G' with $L(G) = L(G')$.

LEMMA 1. *If $G = (V_N, V_T, R, A_n)$ with $V_N = \{A_1, A_2, \cdots, A_n\}$ is a reduced, unambiguous grammar then for no i with $1 \leq i \leq n$ does $A_i \Rightarrow A_i$ hold.*

PROOF. Suppose $A_i \Rightarrow A_i$ for some $i \neq n$. Since G is reduced, $A_i \Rightarrow x$ and $A_n \Rightarrow yA_iz$ for some x,y,z. Now $A_n \Rightarrow yA_iz \Rightarrow yxz$ and $A_n \Rightarrow yA_iz \Rightarrow yA_iz \Rightarrow yxz$ are two derivations of yxz with different lengths. Thus the corresponding leftmost derivations have different lengths, and this contradicts the hypothesis that G is unambiguous. Similarly, $A_n \Rightarrow A_n$ leads to a contradiction by considering $A_n \Rightarrow x$ and $A_n \Rightarrow A_n \Rightarrow x$.

Notation. If $G = (V_N, V_T, R, A)$ is a CF grammar, let $P(G) = \{B \in V_N \mid B \Rightarrow xBy$ for some x,y with $xy \neq \epsilon\}$.

Definition. A CF grammar $G = (V_N, V_T, R, A)$ is called *almost looping* if (i) through (iii) hold:

(i) G is reduced.

(ii) $V_N - \{A\} \subset P(G)$.

(iii) Either $A \in P(G)$ or A occurs just once in the A-derivation of any word $x \in L(G)$.

LEMMA 2. *For any unambiguous reduced CF grammar $G' = (V_N', V_T, R', A)$ there exists an unambiguous almost looping CF grammar $G = (V_N, V_T, R, A)$ with $L(G) = L(G')$.*

PROOF. Let $V_N - \{A\} = \{A_1, A_2, \cdots, A_k\}$ and $A = A_{k+1}$. If $k = 0$ the lemma holds with $G = G'$, since (ii) holds vacuously in this case. If $k > 0$ we define a sequence of unambiguous reduced grammars $G' = G_0, G_1, G_2, \cdots, G_k = G$ as follows. Let $1 \leq i \leq k$ and suppose $G_{i-1} = (V_N^{i-1}, V_T, R^{i-1}, A)$ has already been defined.

Case 1. $A_i \in P(G_{i-1})$. Put G_{i-1}.

Case 2. $A_i \notin P(G_{i-1})$. Consider all the rules of R^{i-1} of the form

(1) $A_i \rightarrow \xi_1, \cdots, A_i \rightarrow \xi_r$.

Then $r \neq 0$ and $\xi_1, \cdots, \xi_r$ do not involve the symbol A_i. We derive R^i from R^{i-1} by removing all rules (1) and replacing any rule of R^{i-1} of the form

(2) $A_j \rightarrow w \quad (j \neq i)$

by $t = r^m$ rules

$$A_j \rightarrow \omega_1, \quad A_j \rightarrow \omega_2, \quad \cdots, \quad A_j \rightarrow \omega_t$$

where m is the number of occurrences of A_i in ω, and $\omega_1, \cdots, \omega_t$ are obtained from ω by replacing each occurrence of A_i in ω by $\xi_1, \xi_2, \cdots, \xi_r$ in all possible ways. Then $G_i = (V_N^i, V_T, R^i, A)$ where $V_N^i = V_N^{i-1} - \{A_i\}$. In both cases it is clear that G_i is unambiguous and reduced and also that $L(G_i) = L(G_{i-1})$. Also $P(G_i) = P(G_{i-1})$ and $(V_N{}^i - P(G_i)) \cap \{A_1, \cdots, A_i\} = \emptyset$. In particular, $V_N{}^k - \{A\} \subset P(G_k)$ and G_k is almost looping.

We now consider $L = \{a^ib^ic^j \mid i,j \geq 1\} \cup \{a^ib^jc^j \mid i,j \geq 1\}$. As is well known, L is CF.

Lemma 3. *If $G = (V_N, V_T, R, A)$ is an almost looping grammar with $L(G) = L = \{a^ib^ic^j \mid i,j \geq 1\} \cup \{a^ib^jc^j \mid i,j \geq 1\}$, then* (i) *through* (iii) *hold:*

(*i*) *Each $B \in V_N - \{A\}$ is of one and only one of the following types:*

Type 1. *There exists an integer $m \geq 1$ and some x and y such that $B \Rightarrow xBy$ with either $xy = a^m$ or $xy = c^m$.*

Type 2. *There exists an integer $n_B \geq 1$ such that $B \Rightarrow a^{n_B}Bb^{n_B}$.*

Type 3. *There exists an integer $n_B \geq 1$ such that $B \Rightarrow b^{n_B}Bc^{n_B}$.*

(*ii*) *A occurs only once in each derivation of any word $x \in L$.*

(*iii*) *There exists a positive number l such that each word $x \in L$ whose derivation contains only A and Type* 1 *nonterminals contains less than l b's.*

Proof of (i) of Lemma 3. We write $a^* = \{a^m \mid m \geq 0\}$, $a^+ = \{a^m \mid m \geq 1\}$, and similarly define b^*, b^+, c^*, and c^+. If $B \in V_N - \{A\}$ then since G is almost looping there are x,y such that $xy \neq \epsilon$ and $B \Rightarrow xBy$. We will show that B is of one of the three types described in the lemma.

We observe first that

$$(1) \quad x,y \in a^* \cup b^* \cup c^*.$$

If this were not the case, and, for example, x contained both a's and b's, e.g. $x = x_1ax_2bx_3$, then L would contain words of the form $x_1ax_2bx_3ax_4$, contrary to the definition of L. In fact since G is reduced and the a's, b's, and c's occur in that order in any word of L, the same is also true for any word derived from a nonterminal B. It follows from this remark that

$$(2) \quad \text{if} \quad x \in c^+ \quad \text{then} \quad y \in c^*,$$

$$(3) \quad \text{if} \quad x \in b^+ \quad \text{then} \quad y \in b^* \cup c^*.$$

We show that

$$(4) \quad xy \notin b^+,$$

$$(5) \quad \text{if} \quad x = a^m \in a^+ \quad \text{then} \quad y \in a^* \cup \{b^m\},$$

$$(6) \quad \text{if} \quad x = b^m \in b^+ \quad \text{then} \quad y \in b^* \cup \{c^m\}.$$

Since G is reduced, $A \Rightarrow \rho B\delta \Rightarrow a^rb^sc^t$ where $r,s,t \geq 1$ and $s = r$ or $s = t$. Since $B \Rightarrow xBy$, we have that

$$A \Rightarrow \rho B\delta \Rightarrow \rho x^qBy^q\delta$$

for any positive integer q. If $xy = b^m \in b^+$ then $A \Rightarrow a^rb^{s+qm}c^t \in L$ for all $q \geq 1$. This is impossible by the definition of L and hence (4) holds. Suppose $x = a^m \in a^+$. If $y = c^n \in c^+$ then $a^{r+qm}b^sc^{t+qn} \in L$ for all q, which is impossible if $m,n \geq 1$. Similarly, if $y = b^n$ and $m \neq n \geq 1$, then $a^{r+qm}b^{s+qn}c^t \in L$ for all $q \geq 1$ and this is impossible. This proves (5). (6) follows by a similar argument.

We now show B is of one of the three types. By (1), $x \in a^* \cup b^* \cup c^*$. If $x = a^m$

($m \geq 1$) then B is of Type 1 or 2 by (5). If $x = b^m$ ($m \geq 1$), then B is of Type 3 by (4) and (6). If $x = c^m$ ($m \geq 1$), then B is of Type 1 by (2). Finally, if $x = \epsilon$, then $y \notin b^+$ by (4) and hence B is of Type 1.

It remains to be shown that the three stated types are mutually exclusive. Suppose B is of both Type 1 and Type 2. Then $B \Rightarrow xBy$, $xy = z^m (m \geq 1)$ where $z = a$ or c, and $B \Rightarrow a^nBb^n$ for some $n \geq 1$. Then

$$B \Rightarrow a^n z^{m_1} B z^{m_2} b^n$$

where $m_1 + m_2 = m$. If $z = c$ this contradicts (1). If $z = a$ and $m_2 \neq 0$, we again contradict (1); if $z = a$ and $m_2 = 0$, we contradict (5). A similar contradiction is deduced if it is assumed B is of Types 1 and 3. Finally, if B is of Type 2 and Type 3, then $B \Rightarrow a^mBb^m$ and $B \Rightarrow b^nBc^n$ where $m,n \geq 1$. Then $B \Rightarrow a^mb^nBc^nb^m$ and this contradicts (1).

Proof of (ii) of Lemma 3. Suppose A occurs at least twice in some derivation of some word $z \in L$. Then by condition (iii) of almost looping $A \Rightarrow xAy$ with $xy \neq \epsilon$. Then according to the proof of (i), A is of Type 1 or 2 or 3, which is impossible. Suppose, for example, $A \Rightarrow xAy$ with $xy = a^m$ and $m \geq 1$. Since abc^2 is in L, $A \Rightarrow abc^2$. Then one obtains $A \Rightarrow a^{m_1}abc^2a^{m_2}$ for some $m_1 + m_2 = m$. This is a contradiction. Or suppose $A \Rightarrow a^{n_B}Ab^{n_B}$ for some $n_B \geq 1$. One obtains $A \Rightarrow a^{n_B}abc^2b^{n_B} \notin L$, again a contradiction. The other cases are treated similarly.

Proof of (iii) of Lemma 3. Suppose words x in L with arbitrarily many b's can be derived from A using only nonterminals of Type 1. Then clearly for some nonterminal B of Type 1, $B \Rightarrow xBy$ with $xy \neq \epsilon$ and xy containing at least one b. Thus B is either of Type 2 or 3, a contradiction.

Theorem. *The CF language* $L = \{a^ib^ic^j \mid i,j \geq 1\} \cup \{a^ib^jc^j \mid i,j \geq 1\}$ *is an inherently ambiguous CF language.*

Proof. Suppose G' is an unambiguous CF grammar with $L(G') = L$. Then there exists a reduced, unambiguous CF grammar G'' with $L(G'') = L$. By Lemma 2 there exists an almost looping, unambiguous CF grammar $G = (V_N, V_T, R, A)$ with $L(G) = L$. Thus Lemma 3 is applicable. Using the notation of that lemma, let $k = \prod n_B$ where the product extends over all $B \in V_N$ of Type 2 or 3. Let $p > l$ be any integer divisible by k. Let $z_1 = a^pb^pc^{2p} \in L$. In an A-derivation of z_1, no nonterminal of Type 3 can occur: suppose $A \Rightarrow \psi B\omega \Rightarrow z_1$ where B is a Type 3 nonterminal. Then $A \Rightarrow \psi B\omega \Rightarrow \psi b^pBc^p\omega \Rightarrow a^pb^{2p}c^{3p}$, a contradiction. Since $p > l$, Type 1 nonterminals alone would not provide enough b's and hence a Type 2 nonterminal B must have been used. Thus $A \Rightarrow \psi B\omega \Rightarrow z_1$ where B is of Type 2, so that $B \Rightarrow a^pBb^p$ holds.

Therefore $A \Rightarrow \psi B\omega \Rightarrow \psi a^pBb^p\omega \Rightarrow a^{2p}b^{2p}c^{2p}$. Thus an A-derivation D_1 of $a^{2p}b^{2p}c^{2p}$ has been found in which no Type 3 but at least one Type 2 nonterminal is used.

Similarly, considering $z_2 = a^{2p}b^pc^p \in L$, an A-derivation D_2 of $a^{2p}b^{2p}c^{2p}$ can be found in which no Type 2 but at least one Type 3 nonterminal is used. The leftmost derivations corresponding to D_1 and D_2 are different. Thus G is ambiguous, which is a contradiction, and the theorem is proved.

REFERENCES

1. Chomsky, N. Formal properties of grammars. In *Handbook of Mathematical Psychology II*, Wiley, New York, 1963, pp. 324–418.
2. Chomsky, N. On certain formal properties of grammars. *Inform. Contr. 1* (1958), 91–112.

. GINSBURG, S. *The Mathematical Theory of Context-Free Languages.* McGraw-Hill, New York, 1966.
. GINSBURG, S., AND ULLIAN, J. S. Ambiguity in context-free languages. *J. ACM 13* (Jan. 1966), 62–89.
. GINSBURG, S. Bounded ALGOL-like languages. *Trans. Amer. Math. Soc. 113* (1964), 333–368.
. PARIKH, R. J. Language-generating devices. Quart. Progr. Rep. No. 60, Res. Lab. Electron., M.I.T., Jan. 1961, 199–212.
. PARIKH, R. J. On context-free languages. *J. ACM 13* (Oct. 1966), 570–581.

ECEIVED FEBRUARY, 1968; REVISED AUGUST, 1968