A Helpful Result for Proving Inherent Ambiguity*

by

WILLIAM OGDEN
Stanford University

A useful theorem [1, Theorem 4.1] of Bar-Hillel, Perles and Shamir gives a necessary condition for a set of words on an alphabet Σ to be a context-free language. We present here a slightly stronger result which can also be used to prove the inherent ambiguity of certain context-free languages.

We assume familiarity with the concepts of context-free languages (hereafter simply called languages), generation trees, and inherent ambiguity. See [4, pp. 62–65] for a summary of terminology and notation.

THEOREM 1. *For each context-free grammar $G = (V, \Sigma, P, S)$ there is an integer k such that for any word ξ in $L(G)$, if any k or more distinct positions in ξ are designated as distinguished, then there is some A in $V \sim \Sigma$ and there are words $\alpha, \beta, \gamma, \delta$, and μ in Σ^* such that:*

(i) $S \Rightarrow^* \alpha A \mu \Rightarrow^* \alpha\beta A \delta\mu \Rightarrow^* \alpha\beta\gamma\delta\mu = \xi$.

(ii) *γ contains at least one of the distinguished positions.*

(iii) *Either α and β both contain distinguished positions, or δ and μ both contain distinguished positions.*

(iv) *$\beta\gamma\delta$ contains at most k distinguished positions.*

Proof. The theorem is established by considering the generation tree corresponding to some derivation of the word ξ. We call a node s a B-node if s has immediate descendants t_1 and t_2 such that both t_1 and t_2 have descendants whose labels occupy distinguished positions in ξ. Let p be the maximum length of the right-hand side of any production in P. We observe that if every path in the derivation tree of ξ contains at most i B-nodes, then ξ contains at most p^i distinguished positions.

Now from those paths in our tree which contain the maximum number of B-nodes and which end at nodes which are distinguished positions in ξ, we select some path $s_0, \cdots, s_{n-1}$. If v is the number of symbols in V, we let $k = p^{2v+3}$, so that the path $s_0, \cdots, s_{n-1}$ will contain at least $2v + 3$ B-nodes. Now choose some final segment $s_m, s_{m+1}, \cdots, s_{n-1}$ of our path which contains exactly $2v + 3$ B-nodes. We divide the B-nodes in this segment into two classes C_L and C_R, so that either C_L or C_R must contain at least $v + 2$

*Preparation of this manuscript supported by Air Force Contract F44620 68-C-0030.

MATHEMATICAL SYSTEMS THEORY, Vol. 2, No. 3
Published by Springer-Verlag New York Inc.

B-nodes. A B-node s_j with $m \leqslant j < n$ is in C_L (C_R) if it has an immediate descendant t which does not have s_{n-1} as a descendant but which does have a distinguished descendant which is to the left (right) of s_{n-1}.

We consider the case where C_L has at least $v + 2$ elements. Find the least integer h such that s_h is in C_L. Now since there are only v different node labels, there must be two integers i and j, where $h < i < j$, such that the nodes s_i and s_j are in C_L and both have the same label A. We get the desired derivation

$$S \Rightarrow^* \alpha A \mu \Rightarrow^* \alpha\beta A \delta\mu \Rightarrow^* \alpha\beta\gamma\delta\mu = \xi$$

by taking γ to be the string of terminal symbols which descend from the node s_j, etc.

Now we check off the other desired properties. γ contains at least one distinguished position since s_{n-1} is a descendant of s_j. β contains a distinguished position since s_i is in C_L. s_i is a descendant of s_h, so, since s_h is in C_L, α also contains a distinguished position. $i > m$, so that the path $s_i, \cdots, s_{n-1}$ contains at most $2v + 3$ B-nodes. Since the path $s_0, \cdots, s_{n-1}$ contains a maximum number of B-nodes, each path in the subtree with vertex s_i contains at most $2v + 3$ B-nodes. Hence $\beta\gamma\delta$ contains at most $p^{2v+3} = k$ distinguished positions.

The case where only C_R has at least $v+2$ elements is, of course, handled analogously.

Applications

By taking all the positions in ξ to be distinguished, we easily obtain the following version of the theorem mentioned at the beginning of this note.

COROLLARY. *For each context-free grammar $G = (V, \Sigma, P, S)$ there is an integer k such that any word ξ in $L(G)$ with $|\xi| \geqslant k$ has a decomposition of the form $\xi = \alpha\beta\gamma\delta\mu$, where $|\beta\delta| > 0$, $|\beta\gamma\delta| \leqslant k$, and for $n \geqslant 0$, $\alpha\beta^n\gamma\delta^n\mu$ is in $L(G)$.*

The following example indicates how our result may also be used to prove inherent ambiguity. Let $M = \{a^iba^{i+1}b \mid i \geqslant 0\}$, $L_0 = abM^*$, and $L_1 = M^*\{a\}^*b$.

THEOREM 2. *The language $L_0 \cup L_1$ is inherently ambiguous.*

Proof. Given any grammar for $L_0 \cup L_1$, we select some integer k_0 with the property ascribed to the k in Theorem 1. Let $p = k_0!$. Our strategy is to find two distinct derivations for the word

$$\xi = aba^2ba^3b \cdots ba^{4p}ba^{4p+1}b.$$

First we look at the word

$$\xi_0 = aba^2ba^3b \cdots ba^{4p-1}ba^pa^pba^{2p+1}b.$$

We designate the first p symbols in the next to last block of a's in ξ as distinguished. Since $p > k_0$, our result gives us an A_0 and α_0, β_0, γ_0, δ_0, and μ_0 with properties (i)–(iv). Since ξ_0 belongs only to L_0 and not to L_1, we see

that properties (i)–(iii) force β_0 to be in the distinguished symbols of ξ_0 and δ_0 to be in the last block of a's in ξ_0. Clearly $\beta_0\gamma_0\delta_0$ cannot be contained entirely within the next to last block of a's in ξ_0, since $S \Rightarrow^* \alpha_0\gamma_0\mu_0$. $\beta_0\gamma_0\delta_0$ cannot contain the third from last b in ξ_0, since β_0 and δ_0 together must contain an even number of b's, and $S \Rightarrow^* \alpha_0\beta_0^2\gamma_0\delta_0^2\mu_0$.

We let $j_0 = |\beta_0| = |\delta_0|$, and by using property (iv), we have $k_0 \geq |\beta_0| = j_0 > 0$. Now for any $q \geq 0$, we have from property (i) that

$$S \Rightarrow^* \alpha_0 A_0 \mu_0 \Rightarrow^* \cdots \Rightarrow^* \alpha_0\beta_0^q A_0 \delta_0^q \mu_0 \Rightarrow^* \alpha_0\beta_0^q\gamma_0\delta_0^q\mu_0.$$

Taking $q_0 = 2k_0!/j_0 + 1$, we get a derivation

$$S \Rightarrow^* \alpha_0\beta_0^{q_0}\gamma_0\delta_0^{q_0}\mu_0 = aba^2b \cdots ba^{4p-1}ba^{p+(q_0-1)j_0}a^pba^{2p+(q_0-1)j_0+1}b.$$

The derived word is just ξ, since $(q_0 - 1)j_0 = 2p$.

Now we look at $\xi_1 = aba^2ba^3b \cdots ba^{4p-2}ba^{2p-1}ba^pa^pba^{4p+1}b$. In this case we designate the last p symbols in the next to last block of a's as distinguished. Again since $p > k_0$, our result gives an A_1 and α_1, β_1, γ_1, δ_1, and μ_1 with properties (i)–(iv). As in the ξ_0 case, we find j_1 and q_1 such that

$$S \Rightarrow^* \alpha_1\beta_1^{q_1}\gamma_1\delta_1^{q_1}\mu_1 = aba^2b \cdots ba^{4p-2}ba^{2p+(q_1-1)j_1-1}ba^pa^{p+(q_1-1)j_1}ba^{4p+1}b = \xi.$$

We see that our two derivations of ξ are incompatible. In the first, the subword $\beta_0^{q_0}\gamma_0\delta_0^{q_0}$ is derived from A_0, while in the second, the subword $\beta_1^{q_1}\gamma_1\delta_1^{q_1}$ is derived from A_1. We chose the distinguished symbols in the two cases in such a way that these two subwords overlap, but neither is contained in the other. So, by a method similar to the one Parikh used [5, Theorem 3], we have established the ambiguity of the grammar.

Our results relate to a couple of open problems which Ginsburg has posed. Theorem 2 yields a negative answer to the following question [3, p. 207]. For each inherently ambiguous language L, is there some bounded regular set $R = \{w_1^* \cdots w_r^*\}$ such that $L \cap R$ is inherently ambiguous? Since the languages L_0 and L_1 both have unambiguous grammars, the inherent ambiguity of $L_0 \cup L_1$ arises from

$$L_0 \cap L_1 = \{aba^2ba^3b \cdots ba^{2n}ba^{2n+1}b \mid n \geq 0\}.$$

But since R is bounded, $(L_0 \cup L_1) \cap R$ contains only a finite number of words in $L_0 \cap L_1$.

Also, by using Theorem 1 in an argument similar to the proof of Theorem 2, we can give a "simple" alternative proof that the language $M_0 \cup M_1$, where $M_0 = \{a^ib^ic^j \mid i, j \geq 1\}$ and $M_1 = \{a^ib^jc^j \mid i, j \geq 1\}$, is inherently ambiguous [3, p. 211]. This proof can be extended to show that the language $(M_0 \cup M_1)^*$ is an example of a language with unbounded degree of inherent ambiguity [2, p. 389]. In fact, we can demonstrate that the problem of determining whether a language has unbounded degree of inherent ambiguity is recursively unsolvable.

For $\xi_0, \cdots, \xi_{n-1}$ in $\{d, e\}^*$, let $N(\xi_0, \cdots, \xi_{n-1})$ be the language generated by the grammar $G = (\{S, d, e, f\}, \{d, e, f\}, P, S)$, where P consists of the productions $S \to \xi_i S d^i e$ and $S \to \xi_i f d^i e$ for each $i < n$. For two sequences

$(\xi_0, \cdots, \xi_{n-1})$ and $(\eta_0, \cdots, \eta_{n-1})$ of words from $\{d, e\}^*$, the question of whether $N(\xi_0, \cdots, \xi_{n-1}) \cap N(\eta_0, \cdots, \eta_{n-1})$ is empty is equivalent to the Post correspondence problem for these two sequences. We can show that the language $(M_0 \cdot N(\xi_0, \cdots, \xi_{n-1}) \cup M_1 \cdot N(\eta_0, \cdots, \eta_{n-1}))^*$ is of unbounded degree of inherent ambiguity if and only if $N(\xi_0, \cdots, \xi_{n-1}) \cap N(\eta_0, \cdots, \eta_{n-1})$ is nonempty, and this establishes the result.

REFERENCES

[1] Y. BAR-HILLEL, M. PERLES and E. SHAMIR, On formal properties of simple phrase structure grammars. *Z. Phonetik Sprachwiss. Kommunikat.* **14** (1961), 143–172.

[2] N. CHOMSKY, Formal properties of grammars. *Handbook of Mathematical Psychology*, Vol. 2 (edited by R. D. Luce, R. Bush, and E. Galanter), John Wiley and Sons, Inc., New York, 1963.

[3] S. GINSBURG, *The Mathematical Theory of Context-Free Languages.* McGraw-Hill Book Company, New York, 1966.

[4] S. GINSBURG and J. S. ULLIAN, Ambiguity in context-free languages. *J. Assoc. Comput. Mach.* **13** (1966), 62–89.

[5] R. J. PARIKH, Language generating devices. *MIT Res. Lab. Electron. Quart. Prog. Rep.* **60** (1961), 199–212.

(Received 10 December 1967)