KYBERNETIKA — VOLUME 9 (1973), NUMBER 4

On Star Height Hierarchies of Context-free Languages

JOZEF GRUSKA

Two definitions of star height of context-free languages are considered. It is shown that the corresponding star height hierarchies of context-free languages are infinite with no gaps and that there is no effective way to determine star height of the language generated by an arbitrary context-free grammar.

1. INTRODUCTION

Two definitions of star height of context-free languages (CFL's) are considered in this paper. They are based on two different characterizations of context-free languages by "substitution expressions" [7] and by "context-free expressions" [5]. It is shown here that it follows easily from the results in [4] that for any of these two definitions of star height and for any integer n there is a linear context-free language star height of which is exactly n. Moreover, it is shown here that there is no effective way to determine star height of the language generated by an arbitrary context-free grammar (CFG). Finally, the two definitions of star height of context-free language are compared and the special case of regular languages is considered.

2. SUBSTITUTION STAR HEIGHT

We start by recalling the main notions and notation from [7] in a little modified form.

If L and L_1 are context-free languages and δ is a symbol, then the operations of substitution $L[\delta \leftarrow L_1]$ and of substitution star $L^{*\delta}$ are defined as follows:

$$L[\delta \leftarrow L_1] = \{w_0 u_1 w_1 \dots u_n w_n;\ u_i \in L_1,\ w_0 \delta w_1 \dots \delta w_n \in L \text{ and } \delta \text{ does not occur in any } w_i\},$$

$$L^{*\delta} = \bigcup_{n \geq 0} (L)_n, \text{ where } (L)_0 = \{\delta\} \text{ and } (L)_{n+1} = (L)_n \cup L[\delta \leftarrow (L)_n].$$

Definition. Let Σ be a finite alphabet. The set $\mathscr{E}_\Sigma$ of substitution expressions E over Σ, and their substitution star heights $\mathrm{sh}_s(E)$, is the smallest set of expressions that can be formed, and their substitution star height defined, by rules 1 and 2 below.

1. If $x \in \Sigma^*$, then $x \in \mathscr{E}_\Sigma$ and $\mathrm{sh}_s(x) = 0$; $\emptyset \in \mathscr{E}_\Sigma$ and $\mathrm{sh}_s(\emptyset) = 0$.*
2. If $E_1 \in \mathscr{E}_\Sigma$, $E_2 \in \mathscr{E}_\Sigma$, $\delta \in \Sigma$, then $(E_1 \cup E_2)$, $E_1[\delta \leftarrow E_2]$ and $E_1^{*\delta}$ are in $\mathscr{E}_\Sigma$ and

$$\mathrm{sh}_s((E_1 \cup E_2)) = \mathrm{sh}_s(E_1[\delta \leftarrow E_2]) = \max\{\mathrm{sh}_s(E_1), \mathrm{sh}_s(E_2)\},$$
$$\mathrm{sh}_s(E_1^{*\delta}) = 1 + \mathrm{sh}_s(E_1).$$

For every $E \in \mathscr{E}_\Sigma$, the language $|E|$ is defined recursicely by

1. $|x| = \{x\}$ if $x \in \Sigma^*$, $|\emptyset| = \emptyset$.
2. If E_1, E_2 are in $\mathscr{E}_\Sigma$, $\delta \in \Sigma$, then

$$|(E_1 \cup E_2)| = |E_1| \cup |E_2|\,; \quad |E_1[\delta \leftarrow E_2]| = |E_1|\,[\delta \leftarrow |E_2|] \text{ and } |E_1^{*\delta}| = |E_1|^{*\delta}.$$

It is shown in [7] that L is a context-free language if and only if there is a substitution expression E such that $|E| = L$.

Substitution star height of a context-free language L, in written $\mathrm{sh}_s(L)$, is defined by $\mathrm{sh}_s(L) = \min\{\mathrm{sh}_s(E);\ |E| = L\}$.

3. DEPTH OF CONTEXT-FREE LANGUAGES

As far as context-free grammars are concerned we use Ginsburg's [3] terminology and notation. If $G = \langle V, \Sigma, P, \sigma\rangle$ is a context-free grammar, then Depth (G) is the maximal integer n such that $V - \Sigma$ contains n distinct nonterminals $A_1, \ldots, A_n$ such that if $1 \leqq i < j \leqq n$, then there are words u, $\bar{u}$, v and $\bar{v}$ such that $A_i \Rightarrow^* uA_jv$ and $A_j \Rightarrow^* \bar{u}A_i\bar{v}$ in G. For a context-free language L let Depth $(L) = \min\{\text{Depth}(G);\ L(G) = L\}$.

4. RESULTS

It is shown in [7] how to construct, given a CFG G, a substitution expression E such that $|E| = L(G)$ and $\mathrm{sh}_s(E) \leqq n$ where n is the number of nonterminals of G. A substitution expression E such that $|E| = L(G)$ can be constructed also in the following way:

Let us say that two nonterminals A and B of G are equivalent if there are words u, v, $\bar{u}$ and $\bar{v}$ such that $A \Rightarrow^* uBv$ and $B \Rightarrow^* \bar{u}A\bar{v}$ in G. Let us now divide context-free equations corresponding to G into several groups in such a way that each group contains equations the left side symbols of which form an equivalence class with

* $\emptyset$ is the symbol for the empty set.

respect to the above defined equivalence on nonterminals of G. Hence, no group has more than Depth (G) equations. Let us now consider separately each group of context-free equations and let us treat those nonterminals of G which are not on a left side of this group of equations as terminals. To any such group of equations and to any of its nonterminals one can construct a substitution expression, star height of which is not more than Depth (G), which represents the language corresponding to the choosen group of equations and to the chosen nonterminal. From such substitution expressions one can get a substitution expression E such that $|E| = L(G)$ using only the operation of substitution. Since substitution does not increase star height, we get that $\mathrm{sh}_s(L) \leqq \mathrm{Depth}(L)$ for any CFL L. On the other hand, it is quite obvious how to construct, given a substitution expression E such that $|E|$ is an infinite language, a CFG G such that $L(G) = |E|$ and $\mathrm{Depth}(G) \leqq \mathrm{sh}_s(E)$. From that the following lemma follows immediately:

Lemma. Depth $(L) = \mathrm{sh}_s(L)$ *for any infinite context-free language L.*

It is shown in [4] that for any integer n there is an infinite linear context-free language $L_n \subset \{0, 1\}^*$ such that Depth $(L_n) = n$. Hence we get .

Theorem 1. *For any integer n there is a linear context-free language* $L_n \subset \{0, 1\}^*$ *such that* $\mathrm{sh}_s(L_n) = n$.

This theorem was also proven in [7] using a result on regular star height hierarchy·

Undecidability of some problems regarding the depth of context-free languages was proven in [6]. From those results and from the Lemma, the following two results follow easily:

Theorem 2. *Let n be an integer. It is undecidable for an arbitrary context-free grammar G whether or not* $\mathrm{sh}_s(L(G)) = n$.

Corollary 3. *There is no effective way to determine* $\mathrm{sh}_s(L(G))$, *given an arbitrary context-free grammar G.*

5. CONTEXT-FREE STAR HEIGHT

As it was shown in [5, 8], context-free languages can be represented also by the so-called "context-free expressions" [5] using union, concatenation and special star operations which are an analog of the star operation for regular sets. Context-free expressions form the base for another definition of the star height of context-free languages.

If L is a language and δ is a symbol, then we define $L^\delta = L^{*\delta}[\delta \leftarrow \emptyset]$.

Definition. Let Σ be a finite alphabet. The set $\bar{\mathscr{E}}_\Sigma$ of context-free expressions E over Σ, and their context-free star height $\mathrm{sh_c}(E)$, is the smallest set of expressions that can be formed, and their context-free star height defined, by rules 1 and 2 below.

1. If $a \in \Sigma \cup \{\varepsilon\}$; then $a \in \bar{\mathscr{E}}_\Sigma$ and $\mathrm{sh_c}(a) = 0$; $\emptyset \in \bar{\mathscr{E}}_\Sigma$ and $\mathrm{sh_c}(\emptyset) = 0$.*
2. If $E_1 \in \bar{\mathscr{E}}_\Sigma$, $E_2 \in \bar{\mathscr{E}}_\Sigma$ and $\delta \in \Sigma$, then $(E_1 \cup E_2)$, $(E_1 \,.\, E_2)$ and $(E_1\delta)$ are in $\bar{\mathscr{E}}_\Sigma$ and

$$\mathrm{sh_c}((E_1 \cup E_2)) = \mathrm{sh_c}((E_1 \,.\, E_2)) = \max\{\mathrm{sh_c}(E_1), \mathrm{sh_c}(E_2)\}\,,$$

$$\mathrm{sh_c}((E_1\delta)) = 1 + \mathrm{sh_c}(E_1)\,.$$

For every $E \in \bar{\mathscr{E}}_\Sigma$, the language $|E|_c$ is defined recursively by

1. If $a \in \Sigma \cup \{\varepsilon\}$, then $|a|_c = \{a\}$; $|\emptyset|_c = \emptyset$.
2. If E_1, E_2 are in $\bar{\mathscr{E}}_\Sigma$ and $\delta \in \Sigma$, then

$$|(E_1 \cup E_2)|_c = |E_1|_c \cup |E_2|_c\,, \quad |(E_1 \,.\, E_2)|_c = |E_1|_c \,.\, |E_2|_c$$

and

$$|(E_1\delta)|_c = |E_1|_c^\delta\,.$$

It is shown in [5, 8] that L is a context-free language if and only if there is a context-free expression E such that $|E|_c = L$.

Context-free star height of a context-free language L, in written $\mathrm{sh_c}(L)$, is defined by $\mathrm{sh_c}(L) = \min\{\mathrm{sh_c}(E), |E|_c = L\}$.

6. RESULTS

For a context-free grammar G let $\mathrm{Var}(G)$ be the number of nonterminals of G and for a context-free language L let $\mathrm{Var}(L) = \min\{\mathrm{Var}(G); L(G) = L\}$.

It is shown in [5] how to construct, given an arbitrary context-free grammar G (a context-free expression E), a context-free expression E (a context-free grammar G) such that $|E|_c = L(G)$. The inspection of these constructions reveals immediately that $\mathrm{Depth}(L) \leqq \mathrm{sh_c}(L) \leqq \mathrm{Var}(L)$ for any context-free language L. It is shown in [4], that for any integer n there is an infinite linear context-free language $L_n \subset \{0, 1\}^*$ such that $\mathrm{Var}(L_n) = \mathrm{Depth}(L_n)$. From that it follows:

Theorem 4. *For any integer n there is an infinite linear context-free language L_n such that* $\mathrm{sh_c}(L_n) = n$.

The last two-results deal with the decision problems concerning context-free star height.

Theorem 5. *Let n be an integer It is unsolvable for an arbitrary context-free grammar G whether or not* $\mathrm{sh_c}(L(G)) = n$.

* The symbol ε denotes the empty word.

Proof. Let $x = (x_1, \ldots, x_m)$, $y = (y_1, \ldots, y_m)$ be arbitrary m-tuples of non-empty words over the alphabet $\{a, b\}$. Let $c, d, e, f, g, h, k, \$$ be symbols not in $\{a, b\}$. Let $L(x)$, $L(x, y)$ and L_s be languages defined by

$$L(x) = \{ba^{i_1} \ldots ba^{i_k} c x_{i_k} \ldots x_{i_1};\ 1 \leqq i_j \leqq m\}\,,$$
$$L(x, y) = L(x)\, c\, L^R(y)\,,$$
$$L_s = \{w_1 c w_2 c w_2^R c w_1^R;\ w_1, w_2 \text{ are in } \{a, b\}^*\}$$

where w^R is the reverse of the word w and for a language L, $L^R = \{w^R; w \in L\}$. By [3], Section 4.2, given x and y, one can effectively construct a context-free grammar $G'_{x,y}$ with the initial symbol σ' and such that $L(G'_{x,y}) = \{a, b, c\}^* - L(x, y) \cap L_s$. Let σ, A, B, ξ be not symbols of $G'_{x,y}$ and let $G_{x,y}$ be the context-free grammar the initial symbol of which is σ and the rules of $G_{x,y}$ are those of $G'_{x,y}$ and, moreover, the rules:

$$\sigma \to A\xi d \mid \xi d\,,$$
$$A \to eA\sigma'\$ \mid eB\xi\$d \mid e\sigma'\$\,,$$
$$B \to eB\xi\$ \mid eA\sigma'\$d \mid e\xi\$\,,$$
$$\xi \to \xi a \mid \xi b \mid \xi c \mid \varepsilon\,.$$

It is easy to verify that if $L(x, y) \cap L_s = \emptyset$, then $L(G_{x,y})$ is exactly the language generated by the grammar

$$\sigma \to Ad\,,$$
$$A \to eA\$ \mid eA\$d \mid Aa \mid Ab \mid Ac \mid \xi$$

and therefore $\mathrm{sh_c}\,(L(G_{x,y})) = 1$.

Let us now assume that $L(x, y) \cap L_s \neq \emptyset$. It is not difficult to verify that if $L(G_{x,y})$ is a sequential language (see [3]), then so is the language L_0 defined by
$L_0 = \{x$; there is a word $y \in \{a, b, c\}^*$ and a word u such that either $x = u\$$ and

$$u\$yd \in L(G_{x,y}) \text{ or } x = ud \text{ and } udyd \in L(G_{x,y})\}\,.$$

However, L_0 is exactly the language generated by the grammar $G''_{x,y}$ which is derived from $G_{x,y}$ by discarding the rule $\sigma \to \xi d$ and by replacing the rule $\sigma \to A\xi d$ with the rule $\sigma \to Ad$. By [2], Lemma 2.1, the language generated by the grammar $G''_{x,y}$ is not sequential. Thus $L(G_{x,y})$ is not a sequential language and therefore $\mathrm{sh_c}\,(L(G_{x,y})) \geqq$ $\geqq 2$ if $L(x, y) \cap L_s \neq \emptyset$. It is the well known result that it is undecidable, given arbitrary x and y, whether or not $L(x, y) \cap L_s = \emptyset$ and therefore we have the theorem for the case $n = 1$.

To show theorem for $n > 1$ we proceed as follows. By Theorem 4, there is an infinite context-free language $L_{n-1} \subset \{g, h\}^*$ such that $\mathrm{sh_c}\,(L_{n-1}) = n - 1$. Let G_{n-1} be a context-free grammar for L_{n-1} with σ_0 being the initial symbol of G_{n-1} and with nonterminals of G_{n-1} distinct from those of $G_{x,y}$. Let $G^0_{x,y}$ be a context-free grammar the rules of which are those of G_{n-1} and of $G_{x,y}$ with the sambol d replaced by the word $d\sigma_0 k$. Since $L(G_{n-1})$ and $L(G_{x,y})$ are languages over disjoint alphabets, one can show on the base of similar arguments as for the case $n = 1$ that $\mathrm{sh}_c(L(G^0_{x,y})) = n$ if and only if $L(x, y) \cap L_s = \emptyset$. Once this is done the theorem for $n > 1$ follows in the same way as for $n = 1$.

Corollary 6. *There is no effective way to determine* $\mathrm{sh}_c(L(G))$ *given an arbitrary context-free grammar G.*

7. RELATIONS BETWEEN STAR HEIGHTS

If L is a context-free language, then it clearly holds $\mathrm{sh}_c(L) \geqq \mathrm{sh}_s(L)$. As we already know, for any integer n there is a context-free language $L_n \subset \{0, 1\}^*$ such that $\mathrm{sh}_c(L_n) = \mathrm{sh}_s(L_n) = n$. On the other hand it can be shown that for any n $\mathrm{sh}_s(L'_n) = 1$ and $\mathrm{sh}_c(L'_n) = n$ for the language $L'_n = \{a^{i_1}ba^{i_2}b \dots a^{i_n}bbaab^{i_n}a \dots b^{i_2}ab^{i_1};\ 1 \leqq i_k,\ 1 \leqq k \leqq n\}$.

If R is a regular set then $\mathrm{sh}_s(R) = 0$ or 1 depending on if R is finite or infinite. It is an open problem whether for any integer n there is a regular set R_n such that $\mathrm{sh}_c(R_n) = n$.

Comparing sh_c with star height sh for regular sets we have that $\mathrm{sh}_c(R) \leqq \mathrm{sh}\, R$ for any regular set R. For any integer n the language R_n generated by the grammar with the rules $\sigma \to \varepsilon$, $\sigma \to \sigma\sigma$, $\sigma \to a\sigma b$, $\sigma \to b\sigma a$, $\sigma \to (a\sigma)^{2n}$, $\sigma \to (b\sigma)^{2n}$ is regular, $\mathrm{sh}_c(R_n) = 1$ and $\mathrm{sh}(R_n) = n$ as it was shown in [1].

(Received January 8, 1973.)

Added in proof: The correction of some proofs will be presented in the next issue.

REFERENCES

[1] F. Dejean, M. P. Schützenberger: On a question of Eggan. Information and Control *9* (1966), 23—25.

[2] S. Ginsburg: Some recursively unsolvable problems in ALGOL-like languages. SDC document SP-670, June 1962.

[3] S. Ginsburg: The mathematical theory of context-free languages, McGraw-Hill, New York 1966.

[4] J. Gruska: A characterization of context-free languages. Information and Control *14* (1970), 152—179.

[5] J. Gruska: A characterization of context-free languages. J. Comput. Sci. *5* (1971), 353—364.

[6] J. Gruska: Complexity and unambiguity of context-free grammars and languages. Information and Control *18* (1971), 502—519.

[7] J. P. McWhirter: Substitution expressions. J. Comput. System. Sci. *5* (1971), 629—637.

[8] M. K. Yntema: Cap expressions for context-free languages. Information and Control *18* (1971), 311—318.

RNDr Jozef Gruska, CSc.: Matematický ústav SAV (Mathematical Institute — Slovak Academy of Sciences), Štefánikova 41, 886 25 Bratislava. Czechoslovakia.