Jozef Gruska

Grammatical levels and subgrammars of context-free grammars

Archivum Mathematicum, Vol. 9 (1973), No. 1, 19--21

Persistent URL: http://dml.cz/dmlcz/104787

ARCH. MATH. 1, SCRIPTA FAC. SCI. NAT. UJEP BRUNENSIS
IX: 19—21 1974

GRAMMATICAL LEVELS AND SUBGRAMMARS OF CONTEXT-FREE GRAMMARS

JOZEF GRUSKA (Bratislava)

(Received May 15, 1972)

1. In this paper the concept of a grammatical level is used to define subgrammars of context-free grammars (CFG's) and this approach is compared to that of Małuszysńki. Moreover, some results concerning the number of subgrammars are derived.

2. As far as CFG's are concerned, we shall use throughout this paper Ginsburg's [1] notation and terminology except that we allow for a CFG to have a finite set of initial symbols.

By a context-free rule p we mean an ordered pair $p = (A, \alpha)$ where A is a symbol and α is a word over a finite alphabet. Instead of $p = (A, \alpha)$ we shall usually write p in the form $A \to \alpha$ and we shall denote $\mathcal{L}(p) = A, \mathcal{R}(p) = \alpha$.

If P is a finite set of context-free rules, then we define

$\mathcal{V}(P)$ to be the smallest alphabet such that $P \subset \mathcal{V}(P) \times \mathcal{V}(P)^*$

$$\mathcal{L}(P) = \{\mathcal{L}(p);\ p \in P\},$$
$$\mathcal{R}(P) = \{\mathcal{R}(p);\ p \in P\},$$
$$\Sigma(P) = \mathcal{V}(P) - \mathcal{L}(P).$$

On a set P of context-free rules the relations $\leqq$, $*\leqq$ and $\equiv$ are defined as follows:

$p_1 \leqq p_2$ if either $\mathcal{L}(p_1) = \mathcal{L}(p_2)$ or the symbol $\mathcal{L}(p_2)$ occurs in the word $\mathcal{R}(p_1)$,
$*\leqq$ is the reflexive and transitive closure of the relation $\leqq$,
$\equiv\ = *\leqq \cap *\leqq^{-1}$ is an equivalence relation on P.

Let $\bar{P} = P/\equiv$ be a quotient set of P relative to the relation $\equiv$. For $P_1, P_2 \in \bar{P}$ let $P_1 *\leqq P_2$ if there are $p_1 \leqq P_1$ and $p_2 \in P_2$ such that $p_1 \leqq^* p_2$. $*\leqq$ is a quasiordering on $\bar{P}$.

3. If $P' \subset P$ for a CFG $G = \langle V, \Sigma, P, S \rangle$, then we define

$$\mathcal{S}_G(P') = \mathcal{L}(P') \cap \mathcal{R}(P - P') \cup S \cap \mathcal{V}(P').$$

Definition 1. Let $G = \langle V, \Sigma, P, S \rangle$ be a CFG with S being a set of initial symbols. For each $P' \subseteqq P$ let $\mathcal{G}(P') = \langle \mathcal{V}(P'), \Sigma(P'), P', \mathcal{S}_G(P') \rangle$. If $P' \in \bar{P}$, then $\mathcal{G}(P')$ is said to be a grammatical level of G. If $\bar{M} \subseteq \bar{P}$, then $\mathcal{H}(M) = \mathcal{G}(\bigcup_{\bar{m} \in \bar{M}} \bar{m})$ is said to be a quasisubgrammar of G. Let $\Gamma(G)$ and $Q(G)$ be the sets of all grammatical levels and quasisubgrammars of G.

If $G_1 = \langle V_1, \Sigma_1, P_1, S_1 \rangle \in \Gamma(G)$ and $G_2 = \langle V_2, \Sigma_2, P_2, S_2 \rangle \in \Gamma(G)$ for a CFG G, then we write $G_1 *\leqq G_2$ if $P_1 *\leqq P_2$.

In the class $Q(G)$ the operations $\vee$ and $\wedge$ are defined as follows. If G_1 and G_2 are CFG's defined as above, then

$$G_1 \vee G_2 = \langle \mathscr{V}(P_1 \cup P_2),\ \Sigma(P_1 \cup P_2),\ P_1 \cup P_2,\ S_G(P_1 \cup P_2)\rangle$$
$$G_1 \wedge G_2 = \langle \mathscr{V}(P_1 \cap P_2),\ \Sigma(P_1 \cap P_2),\ P_1 \cap P_2,\ S_G(P_1 \cap P_2)\rangle$$

Hence $G_0 \in Q(G)$ if, and only if $G_0 = \bigvee_{i=1}^{k} G_i$ for some $G_i \in \Gamma(G)$.

Proposition 1. The algebra $\langle Q(G), \vee, \wedge\rangle$ *is a lattice.*

Proof. The mapping $\mathscr{H}$ is a bijection of the set $2^{\bar{P}}$ onto $Q(G)$ which maps set union $\cup$ in $\vee$ and set intersection $\cap$ in $\wedge$. Thus $\mathscr{H}$ is the isomorphism of $2^{\bar{P}}$ onto $Q(G)$ and therefore $\langle Q(G), \vee, \wedge\rangle$ is the lattice as well as the algebra $\langle 2^{\bar{P}}, \cup, \cap\rangle$.

4. A subset A of a partially ordered set N is said to be a final segment (antichain) of N if $x \in A$, $y \in N$, $x \leqq y$ implies $y \in A$ (if $x \in A$, $y \in A$, $x \leqq y$ implies $x = y$).

If $G = \langle V, \Sigma, P, S\rangle$ is a CFG nad $\bar{M}$ is a final segment of $\bar{P}$, then $\mathscr{H}(\bar{M})$ is said to be a subgrammar of G, Denote by $\Theta(G)$ the set of all subgrammars of G.

Since intersection and union of two final segments of $\bar{P}$ is again a final segment of $\bar{P}$, we have immediately

Proposition 2. *The algebra* $\langle \Theta(G), \vee, \wedge\rangle$ *is the sublattice of the lattice* $\langle Q(G), \vee, \wedge\rangle$.

Theorem 3. $G_0 = \langle V_0, \Sigma_0, P_0, S_0\rangle$ *is a subgrammar of a CFG* $G = \langle V, \Sigma, P, S\rangle$ *if, and only if* $G_0 = \vee\{G';\ G' \in \psi\}$ *where* ψ *is a final segment of* $\Gamma(G)$.

Proof. $\bar{M} \subseteq \bar{P}$ is a final segment of $\bar{P}$ if, and only if $\{\mathscr{G}(\bar{m});\ \bar{m} \in \bar{M}\}$ is a final segment of $\Gamma(G)$. Hence, G_0 is a subgrammar of G if, and only if there is a final segment $\bar{M}$ of $\bar{P}$ such that $G_0 = \mathscr{H}(\bar{M}) = \mathscr{G}(\bigcup_{\bar{m} \in \bar{M}} \bar{m}) = \vee\{\mathscr{G}(\bar{m});\ \bar{m} \in \bar{M}\}$ and $\{\mathscr{G}(\bar{m});\ \bar{m} \in \bar{M}\}$ is a final segment of G. Hence the theorem.

Theorem 4. *The number of subgrammars of a CFG G is equal to the number of antichains of the set* $\Gamma(G)$.

Proof. Let $G = \langle V, \Sigma, P, S\rangle$. If $\bar{E} \subseteq \bar{P}$ ($\bar{A} \subseteq \bar{P}$) is a final segment of $\bar{P}$ (an antichain of $\bar{P}$), then let $\nu(\bar{E})$ be the set of all minimal elements of $\bar{E}$ (let $\mu(\bar{A})$ be the smallest final segment of $\bar{P}$ containing $\bar{A}$). Clearly $\mu\nu$ and $\nu\mu$ are identity mappings on the sets of all final segments and antichains of $\bar{P}$. Hence ν is a bijection of the set of all final segments of $\bar{P}$ onto the set of all antichains of $\bar{P}$. Now the theorem follows from the fact that there exists a bijection of the set of final segments of $\bar{P}$ onto the set $\Theta(G)$ and there is a bijection of the set of antichains of $\bar{P}$ onto the set of antichains of $\Gamma(G)$.

It is evident that the number of subgrammars does not depend on the set of initial symbols.

5. It was shown in Gruska [2] that for any n there is a context-free language $L_n \subset \{a, b\}^*$ such that any CFG generating L_n has at least n grammatical levels. From that and from Theorem 4 it folows immediately.

Theorem 5. *For any integer n there is a context-free language* $L_n \subset \{a, b\}^*$ *such that any CFG generating* L_n *has at least n subgrammars.*

One can even prove a little more, namely, that for any n the language $L'_n = \{a^{k_1}ba^{k_2}b \ldots a^{k_n}bbaab^{k_n}ab^{k_{n-1}}a \ldots b^{k_1}a;\ 0 \leqq k_i < \infty,\ i = 1, 2, \ldots, n\}$ can be generated by a CFG having n subgrammars but not by a CFG with less than n subgrammars.

It was shown in Gruska [3] that it is undecidable for an arbitrary CFG G whether or not the language $L(G)$ can be generated by a CFG having exactly one grammatical level. From that it follows.

Theorem 6. *It is undecidable for an arbitrary CFG G and an arbitrary integer n whether or not there is a CFG generating the language $L(G)$ and having not more than n subgrammars.*

6. In a slightly different way the concept of a grammatical level was defined by Małuszyński [4]. He defines that a quadruple $G' = \langle V', \Sigma', P', S'\rangle$ is a subgrammar of a CFG $G = \langle V, \Sigma, P, S\rangle$ if the following conditions are satisfied.

(1) $V' \subset V$ and if $A \in V' \cap (V - \Sigma)$ and $A \to xby \in P,\ b \in V$, then $b \in V'$ (i. e. V' is a "final segment" of symbols)

(2) $$\Sigma' = \Sigma \cap V'$$

(3) $$P' = \{A \to x;\ A \in V', A \to x \in P\}$$

(4) $$S' = S \cap V' \cup \{a;\, a \in V' \text{ and there exists a rule } A \to xay \in P - P'\}$$

According to this definition it may happen that a subgrammar of a CFG is not a context-free grammar since the set S' may contain terminal symbols. If, however, the condition (4) is modified to have the form

(4a) $$S' = S \cap V' \cup \{B;\, B \in V' - \Sigma' \text{ and there exists a rule } A \to xBy \text{ in } P - P'\}$$

then the two definitions are equivalent. (To be precise we should add the condition $V' = \mathscr{V}(P')$.) To show it we proceed as follows. If the conditions (1) to (3) and (4a) are satisfied with $V' = \mathscr{V}(P')$, then from (1) it follows that P' is a final segment of $\bar{P}$ and (4a) implies $S' = \mathscr{S}_G(P')$ and therefore G' is a subgrammar of G. If, on the other hand, G' is a subgrammar, then P' is a final segment of $\bar{P}$ and therefore the conditions (1) to (3) hold and (4a) follows directly from the definition of $\mathscr{S}_G(P')$.

ACKNOWLEDGEMENT

I should like to thank Professor M. Novotný for several helpful comments on a first draft of this paper.

REFERENCES

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

[2] Gruska J., *Some classifications of context-free languages.* Information and Control 14, 1969 1969, 152–173.

[3] Gruska J., *Complexity and unambiguity of context-free grammars and languages.* Information and Control 18, 1971, 502–519.

[4] Małusziński J., *The lattice of subgrammars of context-free grammars.* Algorytmy, V7, 1970, 21–27. (In Polish).

Jozef Gruska
Matematický ústav SAV
Štefánikova 41
886 25 Bratislava, Czechoslovakia