$LR(k)$ GRAMMARS AND DETERMINISTIC LANGUAGES

BY

DANIEL LEHMANN

ABSTRACT*

A clear proof of the statement that every deterministic language has an LR(1) grammar is given. It uses a definition of *LR* (k) grammars found in Lewis and Stearns and the Ginsburg's simulation of a *pda* by a contex-free grammar.

In this note we give a clear proof of the statement that every deterministic language has an $LR(1)$ grammar. This theorem was first described by Knuth [3] and restated in Hopcroft and Ullman [2] but with a non-satisfactory proof, owing to the fact that the grammar which is claimed to be $LR(k)$ is the grammar that results from Ginsburg's [1] simulation of a pushdown automaton (*pda*) the left-most derivations of which simulate the *pda*, whereas $LR(k)$ grammars are defined by a property of their right-most derivations. Going from left-most to right-most derivations is not easy, and it is not clear how the proof in [2] could be corrected.

We shall use an equivalent definition of $LR(k)$ grammars which is found in Lewis and Stearns [4] and which does not rely on right-most derivations. We include a proof of the equivalence of these two definitions since, as far as we know, no such proof is published.

Our notations are those of Hopcroft and Ullman [2]:

— $a, b, c \cdots$ are terminals.

— $A, B, C \cdots$ are non-terminals.

— $w, x, y \cdots$ are strings of terminals.

— $\alpha, \beta \quad \cdots$ are strings of terminals and non-terminals.

— $\rightarrow$ represents a production.

---

* This work is part of a Master's thesis conducted under the supervision of Professor Eliahu Shamir.

Received September 9, 1971

— $\Rightarrow$ represents an immediate derivation.

— $\overset{*}{\Rightarrow}$ is the transitive closure of $\Rightarrow$.

— $\underset{rt}{\overset{*}{\Rightarrow}}$ represents a right-most derivation.

— $\underset{lf}{\overset{*}{\Rightarrow}}$ represents a left-most derivation.

THEOREM I. *Given a reduced context-free grammar G, the two following properties are equivalent and characterize LR(k) grammars:*

a) *Knuth's definition.*

The grammar G is such that if:

$$S \overset{k}{—}| \underset{rt}{\overset{*}{\Rightarrow}} \alpha A w_3 \Rightarrow \alpha\beta w_3 \; (A \to \beta) \text{ and}$$

$$S \overset{k}{—}| \underset{rt}{\overset{*}{\Rightarrow}} \alpha\beta w'_3 \text{ with } k: w_3 = k: w'_3$$

then the right-most derivation of $\alpha\beta w'_3$ *from* $S\overset{k}{—}|$ *as necessarily the form:*

$$S \overset{k}{—}| \underset{rt}{\overset{*}{\Rightarrow}} \alpha A w'_3 \Rightarrow \alpha\beta w'_3.$$

b) *Lewis and Stearns' definition.*

The grammar G is unambiguous and if:

$$S \overset{k}{—}| \overset{*}{\Rightarrow} w_1 A w_3,\; A \overset{*}{\Rightarrow} w_2 \text{ and}$$

$$S \overset{k}{—}| \overset{*}{\Rightarrow} w_1 w_2 w'_3 \text{ with } k: w_3 = k: w'_3$$

then $S\overset{k}{—}| \overset{*}{\Rightarrow} w_1 A w'_3$.

PROOF. a) $\Rightarrow$ b)

Proposition a) asserts that the left-to-right bottom-up parses of two right canonical forms coincide as long as both strings have the same string of k symbols ahead of the part of the string which is being reduced (the handle in Knuth's terminology). This ensures unambiguity and in the case:

$$S\overset{k}{—}| \overset{*}{\Rightarrow} w_1 A w_3 \overset{*}{\Rightarrow} w_1 w_2 w_3$$

$$S\overset{k}{—}| \overset{*}{\Rightarrow} w_1 w_2 w'_3 \quad k: w_3 = k: w'_3$$

the left-to-right parses of $w_1 w_2 w_3$ and $w_1 w_2 w'_3$ will coincide at least up to the reduction of w_2 to A so:

$$S \overset{k}{—}| \underset{rt}{\overset{*}{\Rightarrow}} \alpha A w'_3 \underset{rt}{\overset{*}{\Rightarrow}} \alpha w_2 w'_3 \underset{rt}{\overset{*}{\Rightarrow}} w_1 w_2 w'_3 \text{ and } S\overset{k}{—}| \overset{*}{\Rightarrow} w_1 A w'_3.$$

b) $\Rightarrow$ a)

$$S \overset{k}{\text{—}|} \underset{rt}{\overset{*}{\Rightarrow}} \alpha A w_3 \Rightarrow \alpha\beta w_3$$

$$S \overset{k}{\text{—}|} \underset{rt}{\overset{*}{\Rightarrow}} \alpha\beta w_3'$$

Since G is reduced there is a w_1 such that $\alpha \overset{*}{\Rightarrow} w_1$ and a w_2 s.t. $\beta \overset{*}{\Rightarrow} w_2$. Therefore:

$$S \overset{k}{\text{—}|} \Rightarrow w_1 A w_3,\ A \overset{*}{\Rightarrow} w_2$$

$$S \overset{k}{\text{—}|} \Rightarrow w_1 w_2 w_3'$$

and from property b) we conclude:

$$S \overset{k}{\text{—}|} \overset{*}{\Rightarrow} w_1 A w_3'.$$

We now know two ways of deriving $w_1 w_2 w_3'$:

$$S \overset{k}{\text{—}|} \underset{rt}{\overset{*}{\Rightarrow}} \alpha\beta w_3' \underset{rt}{\overset{*}{\Rightarrow}} \alpha w_2 w_3' \underset{rt}{\overset{*}{\Rightarrow}} w_1 w_2 w_3'$$

and

$$S \overset{k}{\text{—}|} \overset{*}{\Rightarrow} w_1 A w_3' \Rightarrow w_1 \beta w_3' \overset{*}{\Rightarrow} w_1 w_2 w_3'.$$

Since G is unambiguous they correspond to the same derivation tree and β comes from A:

$$S \overset{k}{\text{—}|} \underset{rt}{\overset{*}{\Rightarrow}} \alpha A w_3' \underset{rt}{\Rightarrow} \alpha\beta w_3' \underset{rt}{\overset{*}{\Rightarrow}} w_1 w_2 w_3'$$

Q.E.D.

THEOREM II. *Every language recognized by a deterministic pushdown automaton (dpda) accepting with empty stack has an LR(0) grammar.*

PROOF. We will use Ginsburg's simulation [1] of a *pda* by a context-free grammar and prove that in the deterministic case it has property b) of Theorem I.

PRELIMINARY REMARK. The simulation uses non-terminals which are triplets $[p, A, q]$ where p and q are states of the *pda* and A is a letter of the stack-alphabet, and $L_{[p\ A\ q]}$, the language generated by $[p, A, q]$, is the set of strings that can be read by the *pda* in state p with only A on the stack that leads it to state q with empty stack. Our remark is that in the deterministic case $L_{[p,A,q]}$ has the prefix property: No strict prefix of a word of $L_{[p\ A\ q]}$ can be in $L_{[p\ A\ q]}$ and, moreover,

$L_{[p,A,q]}$ and $L_{[p,A,r]}$ for $q \neq r$ are prefix-incompatible: no prefix (in the weak sense) of $L_{[p,A,q]}$ can be in $L_{[p,A,r]}$ and conversely.

We proceed to the proof.

1) The grammar is unambiguous.

Consider two left-most derivations of a word x, and the first place they differ:

$$S \underset{lf}{\overset{*}{\Rightarrow}} w[p,A,q]\alpha$$

$$[p,A,q] \to a[r,B_1,s_1]\,[s_1,B_2,s_2]\cdots[s_{n-1},B_n,s_n]$$

$$[p,A,q] \to b[t,C_1,u_1]\,[u_1,C_2,u_2]\cdots[u_{m-1},C_m,u_m]$$

(i) If $a = \varepsilon$, state p and stack-symbol A force an ε-move and then $b = \varepsilon$, and the two derivations correspond to the same move in the *dpda*, so: $r = t$, $n = m$, and $B_i = C_i$, $1 \leqq i \leqq n$. Look at the first i such that $u_i \neq s_i$: $[s_{i-1},B_i,s_i]$ and $[u_{i-1},C_i,u_i] = [s_{i-1},B_i,u_i]$ derive words such that one is the prefix of the other, and this proves: $s_i = u_i$, $1 \leqq i \leqq n$.

(ii) If $a \neq \varepsilon$ then $b = a$, the next letter in x and as in (i) the productions correspond to the same move in the *dpda*.

2) If $S \overset{*}{\Rightarrow} w_1[p,A,q]w_3$, $[p,A,q] \overset{*}{\Rightarrow} w_2$, and $S \overset{*}{\Rightarrow} w_1w_2w_3'$, there is a left-most derivation:

$$S \underset{lf}{\overset{*}{\Rightarrow}} w_1[p,A,q]\,[q,B_1,s_1]\cdots[s_{n-1},B_n,s_n].$$

That is to say, that after reading w_1 and after some ε-moves, the *dpda* is in state p with stack $AB_0\cdots B_n$, then in the derivation of $w_1w_2w_3'$ there is a stage:

$$S \underset{lf}{\overset{*}{\Rightarrow}} w_1[p,A,t]\,[t,B_1,t_1]\cdots[t_{n-1},B_n,t_n].$$

$[p,A,t] \overset{*}{\Rightarrow} w_2'$ such that w_2' is a prefix of w_2 or w_2 is a prefix of w_2'. This shows that $w_2 = w_2'$ and $t = q$ (by our preliminary remark).

$$S \underset{lf}{\overset{*}{\Rightarrow}} w_1[p,A,q]\beta \Rightarrow w_1w_2w_3'$$

Q.E.D.

THEOREM I has two corollaries:

COROLLARY I. *If L is a deterministic language L⊣ has an LR(0) grammar, (since L⊣ is recognized by a dpda accepting with an empty stack).*

COROLLARY II. *If L is a deterministic language, it has an LR(1) grammar.*

REFERENCES

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

2. J. E. Hopcroft, and S. D. Ullman, *Formal languages and their relation to Automata*, Addison-Wesley, 1969.

3. D. E. Knuth, *On the translation of languages from left to right*, Information and Control **8**, (1965), 607–639.

4. R. M. Lewis, II, and R. E. Stearns, *Syntax-directed transductions*, J. Assoc. Comput. Mach. **15** (1968), 465–485.

THE HEBREW UNIVERSITY OF JERUSALEM