JOURNAL OF COMPUTER AND SYSTEM SCIENCES: 1, 317-322 (1967)

Characterizing Derivation Trees of Context-Free Grammars through a Generalization of Finite Automata Theory

J. W. THATCHER

IBM Watson Research Center, Yorktown Heights, New York 10598

Revised November 17, 1967

ABSTRACT

The recognizable sets of value trees (pseudoterms) are shown to be exactly projections of sets of derivation trees of (extended) context-free grammars.

I. INTRODUCTION

If there is a connection between context-free grammars and grammars of natural language, it is undoubtedly, as Chomsky [*1*] proposes, through some stronger concept like that of transformational grammar. In this framework, it is not the context-free language itself that is of interest, but instead, the set of derivation trees (structural descriptions or *P*-markers [*1*]). So from the point of view of transformational grammars, sets of trees are of prime importance as opposed to sets of strings.

These observations should certainly motivate interest in generation systems for sets of trees or their representations instead of the full power of context-free generation (see [*2*]). The purpose of this note is to offer an alternative to the constructions in [*2*]. This alternative, which we call "pseudoautomata theory," may provide a fruitful basis for transformational grammars in the sense that the set of derivation trees of any context-free grammar is shown to be recognizable by a "pseudoautomaton."

In the next section, we describe value trees and equivalent representations, pseudoterms. We provide a brief summary of the definitions and closure theorems for recognizable sets of trees (see [*3*]-[*5*]), and in Section 4 the connection between recognizability and derivation trees is presented.

II. VALUE TREES AND PSEUDOTERMS

We begin by describing the universe of discourse for pseudoautomata, the analog of Σ^* for the conventional theory.[1] A *value tree* on a finite alphabet Σ is a function

[1] X^* is the set of finite strings on X including the empty string λ; $X^+ = X^* - \{\lambda\}$.

571/1/4-1

from a finite closed subset of N_ω (set of strings on ω) into Σ where $U \subseteq N_\omega$ is closed if

(i) $w \in U \wedge w = uv \rightarrow u \in U \qquad (w, u, v \in N_\omega)$

(ii) $wn \in U \wedge m \leqslant n \rightarrow wm \in U \qquad (w \in N_\omega, m, n \in \omega).$

It should be fairly clear that value trees on Σ can be represented graphically by constructing a rooted tree (where the successors of each node are ordered), representing the domain of the function, and labeling the nodes with elements of Σ, representing the values of the function. Thus, in Fig. 1 there are two examples; as a function, the left-hand value tree has domain $\{\lambda, 0, 1, 00, 01\}$ and the value at 00, for example, is A.

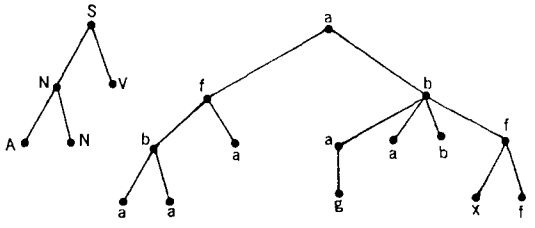

FIG. 1

The definition of value tree and the corresponding pictorial representation provide a good basis for intuition in considering pseudoautomata. The development of the theory, however, is simpler if we consider the familiar linear representation of such trees. For this purpose we define the set $\mathscr{T}_\Sigma$ of *pseudoterms on* Σ as the smallest subset of $(\Sigma \cup \{(,)\})^*$ satisfying:[2]

(i) $\Sigma \subset \mathscr{T}_\Sigma$,

(ii) If $n > 0$ and $f \in \Sigma$ and $t_1, \ldots, t_n \in \mathscr{T}_\Sigma$, then $f(t_1 \cdots t_n) \in \mathscr{T}_\Sigma$.

We will consider value trees and pseudoterms to be equivalent formalizations of the universe of discourse for pseudoautomata theory. The translation between the two is the usual one. By way of example, the value trees of Fig. 1 correspond to the following pseudo-terms:

$$S(N(AN)\,V), \qquad a(f(b(aa)a)b(a(g)abf(xf))).$$

For completeness, we note that this correspondence can be made precise in the following way.

[2] It is assumed that the parentheses are not symbols of Σ.

(i) If $t \in \mathscr{T}_\Sigma$ is *atomic*, i.e., $t = f \in \Sigma$, then the corresponding value tree v_t has domain $\{\lambda\}$ and $v_t(\lambda) = f$.

(ii) If $t = f(t_0 \cdots t_m)$, then v_t has domain $\bigcup_{i \leq m} \{iw \mid w \in \text{domain } (v_{t_i})\} \cup \{\lambda\}$, $v_t(\lambda) = f$, and for $w = iw'$ in the domain of v_t, $v_t(w) = v_{t_i}(w')$.

III. Summary of Generalized Recognizability Concepts

We already have the universe of discourse; namely, the set $\mathscr{T}_\Sigma$ of pseudoterms. Treating the subject in the spirit of Buchi and Wright [6], a *pseudoalgebra on* Σ is a structure $\mathscr{A} = \langle A, \alpha \rangle$ where $\alpha : \Sigma \to A^{A^*}$, i.e., each function $\alpha(f) = \alpha_f$ is defined on A^* taking values in A. A *pseudoautomaton* is a finite pseudoalgebra with the additional restriction that $\alpha_f^{-1}(a) \subseteq A^*$ is regular for each $f \in \Sigma$ and $a \in A$.[3] For $\mathscr{A} = \langle A, \alpha \rangle$, A is the set of *states* and α_f is the *direct transition function* for input f. Analogous to the conventional theory, we can define the *transition function* $\alpha_t : A^* \to A$ for each $t \in \mathscr{T}_\Sigma$ [although here only the value $\alpha_t(\lambda)$ is used] and the definition is simply given by:

$$\alpha_{f(t_1 \ldots t_n)}(a_1 \cdots a_m) = \alpha_f(\alpha_{t_1}(a_1 \cdots a_m) \cdots \alpha_{t_n}(a_1 \cdots a_m)).$$

Then selecting a set $A_F \subseteq A$ of *final states*, the *behavior* of $\mathscr{A}$ relative to A_F is

$$bh_{\mathscr{A}}(A_F) = \{t \mid \alpha_t(\lambda) \in A_F\}.$$

Nondeterministic pseudoautomata are defined in the same way as in the conventional theory; each α_f is a relation in $A^* \times A$. The transition relation α_t for $t \in \mathscr{T}_\Sigma$ is defined by

$$\alpha_{f(t_1 \ldots t_n)}(w, a) \leftrightarrow \exists a_1' \cdots a_n' \left[\bigwedge_i \alpha_{t_i}(w, a_i') \, \alpha_t(a_1' \cdots a_n', a) \right].$$

Then, for a choice of final states, A_F,

$$bh_{\mathscr{A}}(A_F) = \{t \mid \alpha_t(\lambda, a) \wedge a \in A_F\}.$$

A set $U \subseteq \mathscr{T}_\Sigma$ is [*nondeterministically*] *recognizable* if and only if there exists a [nondeterministic] pseudoautomaton $\mathscr{A}$ and set A_F of final states such that $bh_{\mathscr{A}}(A_F) = U$.

To get an intuitive picture of a pseudoautomaton "recognizing" an input pseudoterm, one can describe $\mathscr{A}$ "acting" on the corresponding value three (thus the term

[3] It is assumed that the reader is familiar with the concept of regular set (see [7]).

tree-automaton[4]). The automaton produces a state tree of the same shape (i.e., with the same domain) by assigning an initial state $\alpha_f(\lambda)$ to each terminal node with label f. At any time when states $s_1 \cdots s_k$ have been assigned to all successor nodes of a node labeled f, then that node is assigned the state $\alpha_f(s_1 \cdots s_k)$. The output state or final state is the state assigned to the root of the tree and the tree is accepted if that state is in A_F.

As stated in [8], most of the results of finite automata theory go through for the generalization, pseudoautomata theory. For example:

THEOREM 1. *$U \subseteq \mathscr{T}_\Sigma$ is nondeterministically recognizable iff U is deterministically recognizable.*

By a *projection from* $\mathscr{T}_\Sigma$ to $\mathscr{T}_{\Sigma'}$ we mean a mapping (or relation) $\bar{\pi}$ which is obtained extending $\pi : \Sigma \to \Sigma'$ in the natural way:

$$(0) \qquad \bar{\pi}(f) = \pi(f) \qquad \text{for} \qquad f \in \Sigma,$$

$$(1) \qquad \bar{\pi}(f(t_1 \cdots t_n)) = \pi(f)\,(\bar{\pi}(t_1) \cdots \bar{\pi}(t_n)).$$

Simply stated, $\bar{\pi}(t)$ is the result of replacing each symbol f of t by $\pi(f)$.

THEOREM 2. *The recognizable subsets of $\mathscr{T}_\Sigma$ form a Boolean algebra and projections of recognizable sets are recognizable.*

THEOREM 3. *It is effectively decidable whether $bh_{\mathscr{A}}(A_F) = \varnothing$ for any pseudoautomaton $\mathscr{A}$ and set $A_F \subseteq A$ of final states.*

Some more detail on these and related areas is to be found in [3]-[5]. Work in progress indicates that areas concerned with the semigroup automata, congruences, minimality, and decomposition carry through for the generalization.

IV. Derivation Trees and Recognizability

The connection between generalized recognizability (as applied to pseudoterms) and context-free grammars[5] is quite precisely stated in Propositions 1 and 2 below. To obtain these, we note that derivation trees (P-markers), by Chomsky's definition, *are* value trees. Indeed, we can easily define the set of derivation trees of a context-free grammar in the following way. Let G be such a grammar with terminals Σ, a finite set of nonterminals N, initial set $S_0 \subseteq N$ and a finite set of productions

[4] The concept of three automaton was discovered independently by Doner [3] and Thatcher and Wright [4], the "automata" described here and [5] are slightly more general than those of [3] and [4] in that the symbols of the 'input alphabet' need not have fixed rank.

[5] It is assumed that the reader is acquainted with context-free grammars and languages, see [9] and [10].

$P \subset N \times (\Sigma \cup N)^+$. (We will write $s \rightarrow w$ for $\langle s, w \rangle \in P$.)[6] Denote by D_G^s the set of derivation trees of G from symbol $s \in \Sigma \cup N$. Identifying the value trees with the corresponding pseudo-terms as in Section II, the following is a simultaneous recursive definition of the sets D_G^s:

(i) $s \in D_G^s$ for $s \in \Sigma$,

(ii) if $t_i \in D_G^{s_i}$ $(1 \leqslant i \leqslant n)$ and $s \rightarrow s_1 \cdots s_n$ then $s(t_1 \cdots t_n) \in D_G^s$.

Let $D_G^{S_0} = \bigcup_{s \in S_0} D_G^s$. Then,

PROPOSITION 1. *For any context-free grammar $G = \langle \Sigma, N, S_0, P \rangle$, the set $D_G^{S_0}$ of derivation trees from initial set S_0 to terminal strings is a recognizable subset of $\mathscr{T}_{\Sigma \cup N}$.*

Proof. Given the grammar G, construct the nondeterministic pseudo-automaton $\mathscr{A} = \langle \Sigma \cup N, \alpha \rangle$ on the alphabet $\Sigma \cup N$ where:

(i) $\alpha_s(\lambda, s') \leftrightarrow s = s' \in \Sigma$,

(ii) $\alpha_s(w, s') \leftrightarrow s = s' \wedge s \rightarrow w$.

Two facts must be proved about this construction: (a) that $\mathscr{A}$ *is* a pseudo-automaton; and (b) $\alpha_t(\lambda, s) \leftrightarrow t \in D_G^s$. The first is obvious; the second follows immediately from the definitions by induction. The details will be omitted. From (b), it follows that $D_G^{S_0} = bh_{\mathscr{A}}(S_0)$.

In order to state the converse to Proposition 1, we need to consider a slightly more general concept of context-free grammar. As we have described them, the productions of a context-free grammar comprise a finite subset of $N \times (\Sigma \cup N)^+$. As is well known, nothing new is obtained as far as generated languages are concerned if we allow an infinite set of productions $P \subset N \times (\Sigma \cup N)^+$ with the restriction that for each $s \in N$, $P_s = \{w \mid s \rightarrow w\}$ is regular (or even context-free). Let us call a context-free grammar where regular P_s are allowed, an *extended context-free grammar.* (The need for this extension came to light during a helpful discussion with M. O. Rabin).

Now we can state:

PROPOSITION 2. *Every recognizable subset of $\mathscr{T}_\Sigma$ is a projection of the set of derivations of an extended context-free grammar.*

Proof. Given a Σ-automaton $\mathscr{A} = \langle A, \alpha \rangle$ recognizing U, we define the extended context-free grammar $G = \langle T, N, S_0, P \rangle$ when $T = \{\langle \alpha_f(\lambda), f \rangle \mid f \in \Sigma\}$, $N = A \times \Sigma$, $S_0 = \{\langle a, f \rangle \mid a \in A_F \text{ and } f \in \Sigma\}$ and

$$P = \{\langle a, f \rangle \rightarrow \langle a_1 f_1 \rangle \cdots \langle a_n f_n \rangle \mid \alpha_f(a_1 \cdots a_n) = a, f_i \in \Sigma\}.$$

[6] To simplify the statements below, we have diverged from Chomsky's definition of context-free grammars [2] in two ways. We use an initial set instead of an initial symbol and $\Sigma \cap N = \varnothing$ is not required. This is an insignificant extension. With the usual definition of "languages generated by G," the languages obtained are obviously context-free.

Again, it is easy to check that if t is a derivation in this grammar, the second component of the corresponding tree (the Σ part in $A \times \Sigma$) is an input value tree to $\mathcal{A}$ and the first component is the corresponding state tree. This is a familiar complete state construction and the projection induced by $\pi(a, f) = f$ yields the required recognizable set U.

Putting Propositions 1 and 2 together (using closure of the recognizable sets under projections, Theorem 2), we can state the following characterization of the recognizable subsets of $\mathcal{T}_\Sigma$.

COROLLARY. *A set of value trees (pseudoterms) is recognizable if and only if it is a projection of the set of derivations of some extended context-free grammar.*

The projections are indeed necessary; the set of value trees on {0, 1} with exactly one occurrence of 1 is recognizable, but cannot be obtained directly as a set of derivation trees.

REFERENCES

1. N. CHOMSKY. On the notion of 'rule of grammar'. *in* "Structure of Language and Its Mathematical Aspects" [*Proc. Symposia Appl. Math.*, Vol. XII, pp. 6–24]. American Mathematical Society, Providence, Rhode Island, 1961.
2. S. GINSBURG, AND M. A. HARRISON. *J. Computer Syst. Sci.* 1, 1–23 (1967).
3. J. DONER. Decision Problems of Second-order Logic, (unpublished).
4. THATCHER, J. W., AND J. B. WRIGHT, Generalized finite automata theory with an application to a decision problem of second-order logic, IBM Res. Rept. RC 1713, November 1966: (To appear in *Math. Syst. Theory.*)
5. THATCHER, J. W., "A further generalization of finite automata," IBM Res. Rept. RC 1846, June 1967.
6. J. R. BÜCHI AND J. R. WRIGHT, Mathematical Theory of Automata, notes on material presented by J. R. Büchi and J. B. Wright, Communication Sciences, Course No. 403, Fall 1960. The University of Michigan, Ann Arbor.
7. M. O. RABIN AND D. SCOTT, Finite automata and their decision problems, *IBM J. of Res. Develop.* 3, 114–125 (1959). [Reprinted in "Sequential Machines, Selected Papers". (E. F. Moore, Ed.) Addison-Wesley, Reading, Massachusetts, 1964.
8. J. W. THATCHER, AND J. B. WRIGHT, Abstract 65T-469, *Notices Am. Math. Soc.* 12, 820 (1965).
9. N. CHOMSKY. *Inform. Control* 2, 137–167 (1959).
10. S. GINSBURG. "The Mathematical Theory of Context-Free Languages." McGraw-Hill, New York, 1966.