International Journal of Computer and Information Sciences, Vol. 8, No. 2, 1979

Notes on Canonical Label Languages

Robert L. Cannon, Jr.[1]

Received October 1976; revised March 1978

Derivations by a phrase-structure grammar may be represented by a word over the set of indices for the rules of the grammar. The set of all label words constitutes the label language for the grammar. The canonical label language is the restriction of the label language to the set of canonical derivations of the grammar. Whether a given language may be the label language for any grammar is partially answered by restricting the question to the canonical derivations for regular and contextfree grammars.

KEY WORDS: Grammar; label word; label language; canonical derivation; canonical label language.

1. INTRODUCTION

Placing labels on the productions of a grammar has been found useful by several authors as a means of studying grammars and the languages that they generate. Two approaches have been suggested. One approach has been to consider a language, or *control set*, over the alphabet of labels and to allow a grammar to generate only by a derivation sequence that corresponds to some element of the control set.[1,2] Such an approach has been used to examine the effect of a control set of one type in the Chomsky hierarchy on grammars of various types. A second approach has been to allow a grammar G to generate the language $L(G)$ and to examine the sequence of labels that occur over the set of all possible derivations of elements of $L(G)$.[3–10] This is the label language, or Szilard language, approach.

Clearly, for any grammar G there is an associated label language: the sequence of labels associated with the set of all derivations of elements

[1] Department of Mathematics, Computer Science, and Statistics, University of South Carolina, Columbia, South Carolina.

of $L(G)$. A converse question may be asked. Is every language the label language for some grammar? These notes do not answer that question for arbitrary label languages, but rather for canonical label languages—the label languages associated with the canonical derivations of a grammar. Conditions are described under which, for various types of languages, there exists a grammar whose canonical label language is the given language.

2. DEFINITIONS AND EXAMPLES

A phrase-structure grammar G is a four-tuple $\langle V_N, V_T, P, S\rangle$ such that $V_N \cap V_T = \varnothing$, V_N is defined as a finite set of *nonterminal* symbols with $S \in V_N$, V_T is a set of *terminal symbols*, $V = V_N \cup V_T$ is the alphabet, and the set P of productions satisfies $P \subseteq V^*V_NV^* \times V^*$, where λ is the empty string, V^+ is the closure of V under catenation, and $V^* = V^+ \cup \{\lambda\}$, $P \subseteq V^*V_NV^* \times V^*$ is the set of *productions* of G. A set $\Pi = \{\pi_1, \pi_2, \ldots, \pi_p\}$ in one-to-one correspondence with P is called a set of *labels* for the productions of G. All grammars are assumed to have such a set of labels. For $\pi, \omega, \tau \in V^*$, $\sigma \in V^+$, the string $\pi\sigma\omega$ *derives* the string $\pi\tau\omega$, written $\pi\sigma\omega \Rightarrow \pi\tau\omega$, if there exists a production $\sigma \rightarrow \tau$ in P. The reflexive transitive closure of $\Rightarrow$ is denoted by $\overset{*}{\Rightarrow}$. The *language* $L(G)$ generated by G is defined as $\{x \mid x \in V_T^*, S \overset{*}{\Rightarrow} x\}$.

A grammar is called *regular* iff elements of P are of the form $X \rightarrow aY$, where $X, Y \in V_N$ and $a \in V_T$. A grammar is called *contextfree* iff elements of P are subsets of $V_N \times V^+$. In particular, a contextfree grammar is in Greibach normal form iff productions are of the form $X \rightarrow a\alpha$, $a \in V_T$ and $\alpha \in V_N^*$. When $|\alpha|$ is the length of a string α, a grammar is called *context-sensitive* iff $|\sigma| \leqslant |\tau|$ when $\sigma \rightarrow \tau$ is in P. Regular, contextfree, and context-sensitive grammars generate regular, contextfree, and context-sensitive languages, respectively. All of these definitions follow those of Salomaa.[11]

As first defined by Griffiths[12] for a phrase-structure grammar G, a derivation

$$\alpha_1 \Rightarrow \alpha_2 \Rightarrow \cdots \Rightarrow \alpha_n$$

such that $\alpha_i = \beta_i\sigma_i\gamma_i$, $\alpha_{i+1} = \beta_i\tau_i\gamma_i$, with $\sigma_i \rightarrow \tau_i$ a rule of P, is called a *canonical derivation sequence* iff

$$|\beta_i| < |\beta_{i+1}| + |\sigma_{i+1}|, \qquad 1 \leqslant i \leqslant n-2$$

For a regular or contextfree grammar it is always true that $|\sigma_i| = 1$, so that $|\beta_i| \leqslant |\beta_{i+1}|$, the usual restriction to leftmost derivations for contextfree grammars, and the unique derivation for regular grammars.

If labels are placed on the productions of G, and the labels are recorded from left to right in the sequence of applications of their associated derivations, then a *label word* is formed. Moreover, for a grammar G the *canonical label language* CLL(G) may be defined as the set of all label words associated with canonical derivations of elements of $L(G)$. Canonical derivations for contextfree languages have been studied extensively,[4] Recently, Hart[7] and Cannon[8] have examined canonical derivations for phrase-structure grammars.

Before proceeding to examine the nature of languages that may be a canonical label languages, we should note for several types of grammars their associated canonical label languages.

A contextfree language may be generated by a grammar having a regular, contextfree, or context-sensitive canonical label language. For example, the grammar with production set

$$\{1\colon S \to aSb,\ 2\colon S \to ab\}$$

has canonical label language (1^*2). The grammar with production set

$$\{1\colon S \to aSB,\ 2\colon S \to aB,\ 3\colon B \to b\}$$

has canonical label language $\{1^{n-1}\,23^n \mid n \geqslant 1\}$. Lastly, the grammar with production set

$$\{1\colon S \to SAB,\ 2\colon S \to aB,\ 3\colon BA \to AB,\ 4\colon aA \to aa,\ 5\colon aB \to ab,\ 6\colon bB \to bb\}$$

has canonical label language $\{1^{n-1}\,2343^2\,4 \cdots 3^{n-1}\,456^{n-1} \mid n \geqslant 1\}$. All three grammars generate the contextfree language $\{a^nb^n \mid n \geqslant 1\}$, yet the canonical label languages range from a regular set to a context-sensitive set.

Because of the properties seen in these examples, canonical label languages have been suggested as a measure of the complexity of a grammar. Hart[7] indicates that one of the inadequacies of the canonical label language is that by a trivial reversal of a production of a contextfree grammar the language is reversed, yet the canonical label language for one grammar is contextfree and the other is regular.

Another disadvantage of the canonical label language is that, for a phrase-structure grammar, a canonical label word specifies neither a unique derivation nor the derivation of a unique word in the language. For the grammar with production set

$$\{1\colon S \to aSaSa,\ 2\colon aSa \to aba,\ 3\colon aSa \to aca\}$$

the canonical label word 123 corresponds to two syntactical graphs,[13] as shown in Fig. 1. One graph shows a derivation of *abaca*, the other of *acaba*. This example disproves a conjecture by Hart[7] that a canonical label word may be identified with a unique word in the language generated by G.

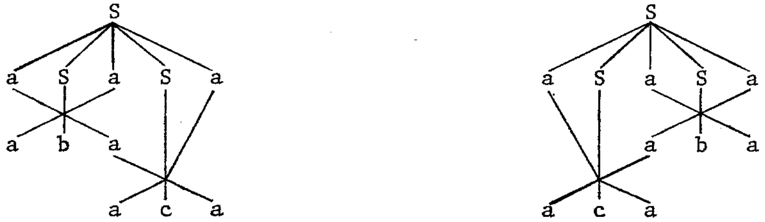

Fig. 1. Derivation of *abaca* and *acaba* associated with the canonical label word 123.

3. LANGUAGES AND ASSOCIATED GRAMMARS

In discussing Szilard (not necessarily canonical) languages, Salomaa[11] notes that it has not been determined whether for a given language L there exists a grammar G such that the Szilard language for G is L. This section answers the question not for arbitrary label languages, but for canonical label languages.

The first problem to be considered here is the existence of a mapping between the canonical label language of some grammar and an arbitrary language.

Proposition 1. For L a regular (contextfree) language there exists a regular (contextfree) grammar G and a length-preserving homomorphism h: CLL(G) $\rightarrow L$ of CLL(G) onto L.

Proof. We need to find a grammar $\hat{G}$ that generates L. For regular L the homomorphism is h: $\pi_i \rightarrow a$ iff π_i is a production in $\hat{G}$ of the form $X \rightarrow aY$ or $X \rightarrow a$. For contextfree L, a grammar $\tilde{G}$ for L in Greibach normal form exists. Again the homomorphism h is h: $\pi_i \rightarrow a$ iff π_i is a production in $\tilde{G}$ of the form $X \rightarrow a\alpha$ where $\alpha \in V_N^*$. With all such grammars there is a one-to-one correspondence between the canonical label words and the elements of L, because at each application of a rewriting rule of $\tilde{G}$ one more terminal symbol is generated.

By Proposition 1 it is seen that there are grammars which behave in a manner similar to any given language, the similarity measured by the number of productions which h projects on a single terminal symbol of L.

Proposition 2. For L a phrase-structure language, there exists a grammar G and a homomorphism h: CLL(G) $\rightarrow L$ of CLL(G) onto L.

Proof. Let L be an arbitrary language and G a grammar for L. For $G = \langle V_N, V_T, P, S\rangle$ define

$$\bar{V}_T = \{\bar{v} \mid v \in V_T\}$$

Let $\hat{G} = \langle \hat{V}_N, V_T, \hat{P}, S\rangle$ such that $\hat{V}_N = V_N \cup \bar{V}_T$, with $\hat{P}$ such that every production of P is replaced by one in $\hat{P}$ such that all elements of V_T are replaced by elements of $\bar{V}_T$, and there are also productions of the form $\bar{a} \rightarrow a$ for all $a \in V_T$. Now a homomorphism h may be defined such that

$$h\colon \pi_i \rightarrow a$$

iff π_i is a production $\bar{a} \rightarrow a$, and otherwise $h\colon \pi_i \rightarrow \lambda$.

In a canonical derivation sequence, each production $\bar{a} \rightarrow a$ is applied when there is no production that can be applied to any symbol to the left of $\bar{a}$. Thus, the labels of the productions of the form $\bar{a} \rightarrow a$ lie in the canonical label word in strictly left-to-right order with respect to the terminals themselves, so that h projects a canonical label word on the word in $L(G)$ that G generated.

Although h maps the canonical label language for G onto L, h has destroyed so much information about the derivation sequences that the elements of CLL(G) and of L bear little resemblance to each other.

Not every language can be the image under a length-preserving homomorphism of the canonical label language of a grammar. Let L be a nonrecursive language, G a grammar, and let h: CLL(G) $\rightarrow L$ be a length-preserving homomorphism of CLL(G) onto L. For $y \in V_T{}^*$, let $J = \{x \mid x \in$ CLL(G), $|x| = |y|\}$. Since J is finite, $K = h(J)$ is finite. Now, $y \in L$ iff $y \in K$. This contradicts the nonrecursiveness of L.

Determining the conditions under which there exists a grammar G such that CLL(G) $= L$ is still the most important question. For regular and contextfree languages the question can be answered partially by the following two propositions.

Proposition 3. Let L be a regular language. There exists a regular grammar G such that CLL(G) $= L$ iff there exists a regular grammar $\hat{G}$ such that $L(\hat{G}) = L$ and no terminal symbol appears in more than one production of $\hat{G}$.

Proof. If such a grammar $\hat{G}$ exists, then productions of $\hat{G}$ can be labeled uniquely by the terminal symbol that is written. Thus, for any word w in $L(\hat{G})$, the corresponding word in CLL($\hat{G}$) is w.

Conversely, let $G = \langle V_N, V_T, P, S\rangle$ be a regular grammar with label set Π and such that CLL(G) $= L$. Construct $\hat{G} = \langle V_N, \Pi, \hat{P}, S\rangle$ such that

$X \to a \in \hat{P}$ iff a: $X \to t \in P$, where $t \in V_T$, and $X \to aY \in P$ iff a: $X \to tY \in P$, $t \in V_T$, $Y \in V_N$. Now $L(\hat{G}) = L$, regular, and, if each production of $\hat{G}$ is labeled by the terminal symbol that it writes, then $L(\hat{G}) = \text{CLL}(\hat{G}) = L$.

There exist regular languages for which no $\hat{G}$ exists. As an example, consider the grammar with production set $\{1\text{: } S \to ATA,\ 2\text{: } T \to UT,\ 3\text{: } U \to a,\ 4\text{: } T \to a,\ 5\text{: } A \to a\}$. This grammar has canonical label language (15)(23)*(45). No regular grammar with a unique production for each of the five labels can generate this language. Thus, this regular language can be the canonical label language for no regular grammar.

Proposition 4. Let L be a contextfree language. There exists a contextfree grammar G such that $\text{CLL}(G) = L$ iff there exists a contextfree grammar $\hat{G}$ in Greibach normal form such that $L(\hat{G}) = L$ and no terminal symbol appears in more than one production of $\hat{G}$.

Proof. If such a $\hat{G}$ exists, then each production of $\hat{G}$ can be labeled by the terminal symbol that it writes. Thus, $\text{CLL}(\hat{G}) = L$.

Conversely, let $G = \langle V_N, V_T, P, S\rangle$ be a contextfree grammar with label set Π and such that $\text{CLL}(G) = L$. For $\beta \in V^+$, let β' be the substring of β with all terminal symbols removed. Construct $\hat{G} = \langle V_N, \Pi, \hat{P}, S\rangle$ such that $X \to a\beta' \in \hat{P}$ iff a: $X \to \beta \in P$ where $\beta \in V^+$ and $\beta' \in V_N^*$. Now $\hat{G}$ is in Greibach normal form, and if each production of $\hat{G}$ is labeled by the production that it writes, then $L(\hat{G}) = \text{CLL}(\hat{G}) = L$.

As an example, for the regular language (15)(23)*(45) the grammar $\hat{G}$ in Greibach normal form has productions

$$\{1\text{: } S \to 1ATA,\ \ 2\text{: } T \to 2UT,\ \ 3\text{: } U \to 3,\ \ 4\text{: } T \to 4,\ \ 5\text{: } A \to 5\}$$

Now $L(\hat{G}) = \text{CLL}(\hat{G}) = (15)(23)^*(45)$. The language $\{1^n 2^n \mid n \geqslant 1\}$ cannot, however, be the canonical label language for any contextfree grammar.

There exist grammars with a contextfree canonical label language, which language cannot, however, be the canonical label language for a contextfree grammar. As an example, the contextfree language

$$\{1\ 2^n\ 3\ 4^{n+1}3 \mid n \geqslant 0\} \tag{1}$$

is the canonical language for the grammar with production set

$$\{1\text{: } S \to AXAa,\ 2\text{: } XA \to XXA,\ 3\text{: } Aa \to aa,\ 4\text{: } Xa \to aa\}$$

No grammar in Greibach normal form can generate this language (1), since there cannot be in the grammar productions of all three forms $X \to 2\alpha$, $X \to 3\beta$, $X \to 4\gamma$, $X \in V_N$, where α, β, $\gamma \in V_N^*$. Thus, the class of context free languages which may be the canonical label language for a phrase-structure grammar includes as a proper subset the class of contextfree languages that may be the canonical label language for a contextfree grammar.

4. THE GENERAL CASE

A characterization of the conditions under which an arbitrary language can be the canonical label language for some grammar is still an open question. It is immediate that not every language can be a canonical label language. For example, the language given by the regular expression 11^* cannot be, nor can the language $\{1^n2^n \mid n \geqslant 1\}$. In the former case it should be apparent that production a cannot both initiate and terminate a canonical derivation sequence. The latter case is shown by the following:

Proposition 5. The language $L = \{1^n2^n \mid n \geqslant 1\}$ cannot be the canonical label language for any phrase-structure grammar.

Proof. Let $G = \langle V_N, V_T, P, S\rangle$ be a phrase-structure grammar with $\mathrm{CLL}(G) = L$. Let $1^k2^k \in L$ where $k \geqslant 1$. Let $w \in L(G)$ be derived by the canonical derivation 1^k2^k.

By the definition of a canonical derivation for a phrase-structure grammar, production 1 must be applied k times before production 2 is applied. Since production 1 is the initial production in the derivation sequence, it must be of the form $S \rightarrow \alpha S\beta$ where α, $\beta \in V^*$. Clearly, S must appear on the left because the initial string of the derivation consists only of S. S must appear on the right because production 1 must be repeated $k - 1$ more times. Note that S might appear as a symbol in α or β.

Assume that $S \overset{*}{\Rightarrow} \gamma S\delta$ where γ, $\delta \in V^*$ in the first k steps the canonical derivation 1^k2^k. Production 2 must have S on its left side in order that a word in $L(G)$ be written by k steps of the derivation sequence. If production 2 is of the form $\eta S\zeta \rightarrow \sigma S\tau$ where η, ζ, σ, $\tau \in V^*$, then an S remains in the sentential form after k applications of production 2. Alternatively, production 2 may be of the form $\eta S\zeta \rightarrow \rho$ where η, ζ, $\rho \in V^*$ and S is not a symbol in ρ. If production 1 had $j + 1$ copies of S on its right side, then there would be $kj + 1$ copies of S in the sentential form $\gamma S\delta$. No sequence of k applications of production 2, however, can rewrite $kj + 1$ copies of the symbol S such that a word in $L(G)$ can be written.

There is still no characterization of languages which can be the Szilard language or the canonical label language for an arbitrary grammar. An upper bound is provided by the observation that all Szilard and canonical label languages are type 1.[2,7] These notes characterize the type 3 and type 2 languages, which may be the canonical label language for grammars of the same respective type. Although the type of the canonical label languages for a grammar can be used as a measure of the complexity of a derivation, it is, however, of little consequence in determining the complexity of the language generated by the grammar.

REFERENCES

1. S. Ginsburg and E. H. Spanier, "Control sets on grammars," *Math. Syst. Theory* **2**:159–177 (1968).
2. A. Salomaa, "On grammars with restricted use of productions," *Ann. Acad. Sci. Fenn., Ser. AI* **454** (1969).
3. T. Jarvi, "On control sets induced by grammars," *Ann. Acad. Sci. Fenn. Ser. AI***480** (1970).
4. E. Moriya, "Associate languages and derivational complexity of formal grammars and languages," *Inf. Control* **22**:139–162 (1973).
5. M. Penttonen, "On derivation languages corresponding to context-free grammars," *Acta Inf.* **3**:285–291 (1974).
6. A. C. Fleck, "An analysis of grammars by their derivation sets," *Inf. Control* **24**:389–398 (1974).
7. J. Hart, "Label Languages and Control Sets of Canonical Phrase Structure Derivations, Technical Report No. 15, University of Kentucky, Lexington (1975).
8. R. L. Cannon, "Phrase structure grammars generating context-free languages," *Inf. Control* **29**:252–267 (1976).
9. H. P. Kriegel and H. A. Maurer, "Formal translations and Szilard languages," *Inf. Control* **30**:187–198 (1976).
10. J. M. Hart, "The derivation language of a phrase structure grammar," *J. Comput. Syst. Sci.* **12**:64–79 (1976).
11. A. Salomaa, *Formal Languages* (Academic Press, New York, 1973).
12. T. V. Griffiths, "Some remarks on derivations in general rewriting systems," *Inf. Control* **12**:27–54 (1968).
13. J. Loeckx, "The parsing for general phrase structure grammars," *Inf. Control* **16**:443–464 (1970).