A NOTE ON REGULAR CANONICAL SYSTEMS

by

A. Ehrenfeucht*

and

G. Rozenberg**

All correspondence to the second author.

*University of Colorado, Computer Science Department, Boulder, CO 80309, USA.

**Institute of Applied Mathematics and Computer Science, University of Leiden, Leiden, The Netherlands.

ABSTRACT.

We give a rather elementary proof of the Büchi theorem (in a somewhat extended version) that regular canonical systems generate regular languages only.

INTRODUCTION.

Regular canonical systems form a special case of canonical systems of Post ([P]) and were considered in depth for the first time by R. Büchi ([B]).

These systems play a basic role in the classical formal language theory (see, e.g., [S] and [G]) and it was demonstrated recently that they play a crucial role in a unified theory of grammars and automata (see [R]). The main result of [B] says that regular canonical systems generate only (and all) regular languages. The original proof of this result by Büchi is really involved (a clear presentation of this proof is given in [S]) - a simpler proof relying on a couple of results concerning push-down automata is given in [G] (actually a somewhat more general theorem is proved in [G]).

We provide a simple proof of Büchi theorem (in the more general version of Greibach). Our proof is simpler and more elementary than the proof in [G] in that it gives a direct construction which for a given regular canonical system yields a finite automaton defining the same language.

PRELIMINARIES.

We assume that the reader is familiar with the basic theory of regular languages (see, e.g., [S]). Perhaps the following notational matters should be pointed out:

- for sets A and B, $A-B$ denotes the set theoretical difference of A and B,
- Λ denotes the empty word,
- for a word x over an alphabet Σ, $pref(x) = \{y \in \Sigma^* : yz = x$ for some $z \in \Sigma^*\}$,
- for words x, y such that y is a prefix of x, $y \backslash x$ denotes the word z such that $yz = x$,
- for languages K and L over an alphabet Σ, the ***right quotient*** of K by L, denoted K/L, is equal to $\{x \in \Sigma^*$: there exists a $y \in L$ such that $xy \in K\}$.

We recall now the definition of a regular canonical system (see, e.g., [B], [G], [S]). Our formulation is the closest to this from [G], however we use the *right* version of regular canonical systems, i.e., the rewriting takes place at the right (rather than the left) end of a string - this does not make any difference as far as the class of defined languages is concerned.

Definition. (1) A *(right-)regular canonical system*, abbreviated rc *system*, is a quintuple $G = (\Sigma, \Delta, U, V, P)$, where Σ is a finite alphabet - the *total alphabet (of G)*, $\Delta \subseteq \Sigma$ is the *terminal alphabet (of G)*, $U \subseteq \Sigma^*$ is the *axiom set (of G)*, $V \subseteq \Sigma^*$ is the *end set (of G)* and $P \subseteq \Sigma^* \times \Sigma^*$ is a finite set of *productions (of G)*.

(2) For G as above and $x, y \in \Sigma^*$ we write $x \underset{G}{\Longrightarrow} y$ if $x = x'\alpha$ and $y = x'\beta$ for some $x' \in \Sigma^*$ and $(\alpha, \beta) \in P$ (we say then that x *directly derives* y *using* (α,β)).

(3) The transitive and reflexive closure of $\underset{G}{\Longrightarrow}$ is denoted by $\overset{*}{\underset{G}{\Longrightarrow}}$.

(4) The *language generated by G*, denoted $L_g(G)$, is defined by

$L_g(G) = \{x \in \Delta^*$: for some $u \in U$ and $v \in V$, $u \overset{*}{\underset{G}{\Longrightarrow}} xv\}$ and the *language accepted by G*, denoted $L_a(G)$, is defined by

$$L_a(G) = \{x \in \Delta^*\text{: for some } v \in V \text{ and } u \in U,\ xv \overset{*}{\underset{G}{\Longrightarrow}} u\}. \blacksquare$$

THE MAIN RESULT.

We will show that for an rc system both, its generated and its accepted language are regular by proving this result for a special subclass of rc systems - called elementary - and then reducing the general regularity problem to the regularity problem for elementary rc systems.

Definition. A *regular canonical system* $G = (\Sigma, \Delta, U, V, P)$ is called *elementary*, abbreviated erc *system*, if $\# U = 1$, $V = \{\Lambda\}$ and $\Sigma = \Delta$. ∎

Remark. (1) Since for an erc system specifying Σ and V is superficial, we will specify an erc system in the form $G = (\Delta, u_{in}, P)$, where $U = \{u_{in}\}$. (2) Note that for an erc system G as above we have $L_g(G) = \{x \in \Delta^* : u_{in} \overset{*}{\underset{G}{\Longrightarrow}} x\}$ and $L_a(G) = \{x \in \Delta^* : x \overset{*}{\underset{G}{\Longrightarrow}} u_{in}\}$. ∎

In order to prove that $L_g(G)$ is regular for an arbitrary erc system G we give a construction which (for a given erc system G) yields a transition system accepting $L_g(G)$. A *transition system* is like (the graph of) a finite automaton except that transitions can be made on words, rather than on letters only, including the empty word. A transition system will be specified in the form $\mathbf{A} = (\Theta, Q, F, q_{in}, \delta)$, where Θ is its alphabet, Q its set of states, F its set of final states, q_{in} its initial state and δ its control consisting of a finite number of *transitions*, each of the form (q, x, q'), where $q, q' \in Q$ and $x \in \Sigma^*$ (in the graph of $\mathbf{A}$ (q, x, q') corresponds to the directed arc from q to q' labelled by x). $L(\mathbf{A})$ denotes the language accepted by $\mathbf{A}$.

Here is the above mentioned construction.

CONSTRUCTION.

Let $G = (\Delta, u_{in}, P)$ be an erc system.

Let $\mathbf{A}_G = (\Theta, Q, F, q_{in}, \delta)$ be the transition system such that $\Theta = \Delta$, $Q = P \cup \{u_{in}\} \cup \{f\}$, where $f \notin P \cup \{u_{in}\}$, $q_{in} = u_{in}$ and δ is defined as follows.

(1) For $\pi_1 = (\alpha_1, \beta_1)$, $\pi_2 = (\alpha_2, \beta_2) \in P$ and $x \in \Delta^+$,

$(\pi, x, \pi_2) \in \delta$ whenever $x \in \mathit{pref}(\beta_1)$ and $x \backslash \beta_1 \overset{*}{\underset{G}{\Longrightarrow}} \alpha_2$.

(2) For $\pi = (\alpha, \beta) \in P$ and $x \in \Delta^+$,

$(\pi, x, f) \in \delta$ whenever $x \in \mathit{pref}(\beta)$ and $x \backslash \beta \overset{*}{\underset{G}{\Longrightarrow}} \Lambda$.

(3) For $\pi = (\alpha, \beta) \in P$ and $x \in \Delta^*$,

$(u_{in}, x, \pi) \in \delta$ whenever $x \in \mathit{pref}(u_{in})$ and $x \backslash u_{in} \overset{*}{\underset{G}{\Longrightarrow}} \alpha$.

(4) For $x \in \Delta^*$,

$(u_{in}, x, f) \in \delta$ whenever $x \in \mathit{pref}(u_{in})$ and $x \backslash u_{in} \overset{*}{\underset{G}{\Longrightarrow}} \Lambda$.

(5) δ contains only transitions specified under (1) through (4) above. ∎

Remark. (1) Requiring that in (1) and (2) above $x \in \Delta^+$ rather than $x \in \Delta^*$ is not necessary however it makes the proof of the next lemma easier (more intuitive).

(2) The effectiveness of the above construction relies on the effectiveness of the $\overset{*}{\underset{G}{\Longrightarrow}}$ relation. However it can be proved that for each rc system the $\overset{*}{\underset{G}{\Longrightarrow}}$ is recursive and so the above construction is constructive. ∎

We will prove now that for G and $\mathbf{A}_G$ as in the above construction the language accepted by $\mathbf{A}_G$ equals the language generated by G. (In what follows a *derivation* of a word $x \in \Delta^*$ in G is a sequence of words $u_0, \ldots, u_n$ - called the *trace* of the derivation - such that $n \geq 0$, $u_0 = u_{in}$ and $u_n = x$ *together* with a sequence of productions $\pi_1, \ldots, \pi_n$ such that, for $1 \leq i \leq n$, u_{i-1} directly derives u_i using π_i; we say that π_i *is used in the i-th step* of the derivation.)

Lemma 1. Let G and $\mathbf{A}_G$ be as in the above construction. Then $L(\mathbf{A}_G) = L_g(G)$.

Proof.

(1) $L(\mathbf{A}_G) \subseteq L_g(G)$.

Consider $x \in L(\mathbf{A}_G)$ and a successful path τ_x in $\mathbf{A}_G$ yielding the acceptance of x.

(i) If τ_x is of the form (the initial state is indicated by a short double arrow and the final state by a double circle)

Figure 1.

then clearly $u_{in} \overset{*}{\underset{G}{\Longrightarrow}} x$ and so $x \in L_g(G)$.

(ii) If τ_x is of the form

Figure 2.

where $n \geq 1$, $x = x_1 \cdots x_n x_{n+1}$, $u_0 = u_{in}$ and $u_i = (\alpha_i\ , \beta_i)$ for $1 \leq i \leq n$, then by the definition of $\mathbf{A}_G$,

$u_{in} = x_1 z_1$ for some $z_1 \in \Delta^*$ such that $z_1 \overset{*}{\underset{G}{\Longrightarrow}} \alpha_1$,

for $1 \leq i \leq n-1$, $\beta_i = x_{i+1} z_{i+1}$ for some $z_{i+1} \in \Delta^*$ such that $z_{i+1} \overset{*}{\underset{G}{\Longrightarrow}} \alpha_{i+1}$,

and $\beta_n = x_{n+1} z_{n+1}$ for some $z_{n+1} \in \Delta^*$ such that $z_{n+1} \overset{*}{\underset{G}{\Longrightarrow}} \Lambda$.

Consequently in G we have

$$u_{in} = x_1 z_1 \overset{*}{\underset{G}{\Longrightarrow}} x_1 \alpha_1 \underset{G}{\Longrightarrow} x_1 \beta_1 = x_1 x_2 z_2 \overset{*}{\underset{G}{\Longrightarrow}} x_1 x_2 \alpha_2 \underset{G}{\Longrightarrow} x_1 x_2 \beta_2 \overset{*}{\underset{G}{\Longrightarrow}} \cdots$$

$$\overset{*}{\underset{G}{\Longrightarrow}} x_1 x_2 \cdots x_n \alpha_n \underset{G}{\Longrightarrow} x_1 x_2 \cdots x_n \beta_n = x_1 x_2 \cdots x_n x_{n+1} z_{n+1} \overset{*}{\underset{G}{\Longrightarrow}} x_1 x_2 \cdots x_{n+1}.$$

Thus $x = x_1 \cdots x_n x_{n+1} \in L_g(G)$.

(2) $L_g(G) \subseteq L(\mathbf{A}_G)$.

consider $x \in L_g(G)$.

If $x = \Lambda$, then $u_{in} \overset{*}{\underset{G}{\Longrightarrow}} \Lambda$ and so, by the definition of $\mathbf{A}_G$, $(u_{in}, \Lambda, f) \in \delta$ and consequently $\Lambda \in L(\mathbf{A}_G)$.

If $x = u_{in}$, then, by the definition of $\mathbf{A}_G$, $(u_{in}, u_{in}, f) \in \delta$ and consequently $u_{in} \in L(\mathbf{A}_G)$.

So we may assume that $x \not\in \{u_{in}, \Lambda\}$. Let D_x be a derivation of x in G with the trace $w_0, w_1, \ldots, w_n$, where $n \geq 1$, $w_0 = u_{in}$ and $w_n = x$, and let, for $1 \leq i \leq n$, $\pi_i = (\alpha_i, \beta_i)$ be the production used in the i-th step. Let, for $0 \leq i \leq n$, $w_i = a_{i,1} \cdots a_{i,m_i}$, where $a_{i,j} \in \Delta$ for $0 \leq i \leq n$ and $1 \leq j \leq m_i$. Also, for $1 \leq i \leq n$, $rem(w_i) = |w_i / \beta_i|$ and $rem(w_0) = 0$.

We will prove that $x \in L(\mathbf{A}_G)$ by demonstrating a "parsing strategy" for x in G which is "realizable" (i.e., corresponds to a successful path) in $\mathbf{A}_G$.

We write w_n in the form $w_n = t_0 t_1 \cdots t_n$ where, for each $0 \leq j \leq n$, t_j is defined as follows:

for each $1 \leq k \leq m_n$,

a_k belongs to t_j if and only if $rem(j) < k \leq rem(l)$ for each $j \leq l \leq m$.

Now let $M = \{j \in \{0, \ldots, n\} : t_j \neq \Lambda)$ and let $(j_1, j_2, \ldots, j_r)$ be the ordering of M such that $0 \leq j_1 < j_2 < \cdots < j_r \leq n$; hence $w_n = t_{j_1} t_{j_2} \cdots t_{j_n}$.

We consider separately two cases.

(i) $j_1 = 0$.

Then $w_0 = u_{in} = t_{j_1} y_1$ for some $y_1 \in \Delta^*$ such that $y_1 \overset{*}{\underset{G}{\Longrightarrow}} \alpha_{j_2}$,

$w_{j_2} = t_{j_1} t_{j_2} y_2$ for some $y_2 \in \Delta^*$ such that $\beta_{j_2} = t_{j_2} y_2$ and $y_2 \overset{*}{\underset{G}{\Longrightarrow}} \alpha_{j_3}$,

$\vdots \qquad \vdots$

$\vdots \qquad\qquad\qquad \vdots$

$w_{j_r} = t_{j_1} t_{j_2} \cdots t_{j_r} y_r$ for some $y_r \in \Delta^*$ such that $\beta_{j_r} = t_{j_r} y_r$ and $y_r \overset{*}{\underset{G}{\Longrightarrow}} \Lambda$.

This may be illustrated as follows:

Figure 3.

Consequently

Figure 4.

is a successful path in $\mathbf{A}_G$.

Thus $t_{j_1} t_{j_2} \cdots t_{j_r} = w_n = x \in L(\mathbf{A}_G)$.

(ii) $j_1 > 0$.

Then we notice that $u_{in} \overset{*}{\underset{G}{\Longrightarrow}} \alpha_{j_1}$ and so $(u_{in}, \Lambda, \pi_{j_1}) \in \delta$.

Now proceeding analogously to (i) above (but "restarting the derivation" from w_{j_1}) we complete this transition to a successful path in $\mathbf{A}_G$:

Figure 5.

Hence $t_{j_1} t_{j_2} \cdots t_{j_r} = w_n = x \in L(\mathbf{A}_G)$.

Consequently $L_g(G) \subseteq L(\mathbf{A}_G)$.

The lemma follows from (1) and (2). ■

Now the regularity of $L_g(G)$ and $L_a(G)$ is proved as follows.

Theorem 1. If G is an erc system then both $L_g(G)$ and $L_a(G)$ are regular.

Proof.

Let $G = (\Delta, u_{in}, P)$.

The regularity of $L_g(G)$ follows directly from Lemma 1.

To see that $L_a(G)$ is regular we notice that for the erc system $G_{inv} = (\Delta, u_{in}, \{(\beta,\alpha) : (\alpha,\beta) \in P\})$ we have $L_g(G_{inv}) = L_a(G)$. Hence, again, the regularity of $L_a(G)$ follows from Lemma 1. ∎

Now in order to extend our results to arbitrary rc systems we note the following - we discuss the generated languages, the situation for accepted languages is analogous.

(1) If $G = (\Sigma, \Delta, U, V, P)$ is an rc system such that $G' = (\Delta, \Delta, U, V, P)$ is elementary, then $L_g(G) = L_g(G') \cap \Delta^*$ and so, by Theorem 1, and the closure of regular languages under intersections, $L_a(G)$ is regular.

(2) If $G = (\Sigma, \Delta, U, V, P)$ is an rc system such that $V = \{\Lambda\}$ and U is regular then we proceed as follows. Let $H = (\Theta, \Delta, R, S)$ be a right-linear grammar (where Θ is its total alphabet, Δ its terminal alphabet, R its set of productions and S its axiom) such that $L(H) = U$; clearly we may assume that $(\Theta - \Delta) \cap (\Sigma - \Delta) = \emptyset$. Then let $G' = (\Sigma', \Delta, \{S\}, V, P')$ be the rc system such that $\Sigma' = \Sigma \cup (\Theta - \Delta)$ and $P' = P \cup R$. Obviously $L_g(G) = L_g(G')$ and so, by (1) above, $L_g(G)$ is regular.

(3) Finally if $G = (\Sigma, \Delta, U, V, P)$ is an rc system such that U and V are regular then we consider the rc system $G' = (\Sigma, \Delta, U, \{\Lambda\}, P)$. Clearly $L_g(G) = L_g(G')/V$ and so, by (2) above and because the family of regular languages is closed under right derivatives, $L_g(G)$ is regular.

Hence as a corollary of Theorem 1 (and simple closure properties of regular languages) we get the following result which generalizes the original theorem of Büchi from [B] (in the version as below the theorem was first proved in [G] and

[K]).

Corollary 1. If $G = (\Sigma, \Delta, U, V, P)$ is an rc system such that U and V are regular languages, then both $L_g(G)$ and $L_a(G)$ are regular. ∎

Remark. (1) Clearly, given an rc system G, we could have given directly the construction of a transition system $\mathbf{A}'_G$ accepting $L_g(G)$. However we felt that going first through elementary rc systems and then reducing the problem to the general case makes the whole idea (behind the main construction) more transparent. (2) Note that Corollary 1 holds for arbitrary V (the class of regular languages is closed with respect to right quotients with arbitrary languages) - the regularity of V is needed to make this result effective. ∎

ACKNOWLEDGEMENTS.

The authors gratefully acknowledge the support of NSF grant MCS 83-05245. The authors are obliged to H.J. Hoogeboom and H.C.M. Kleijn for useful comments concerning the first version of this paper.

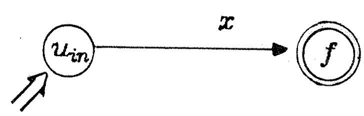

Figure 1.

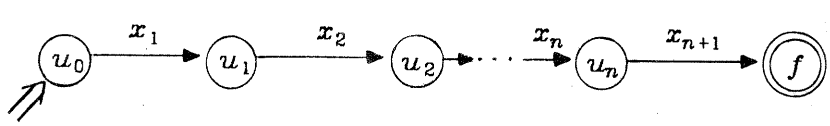

Figure 2.

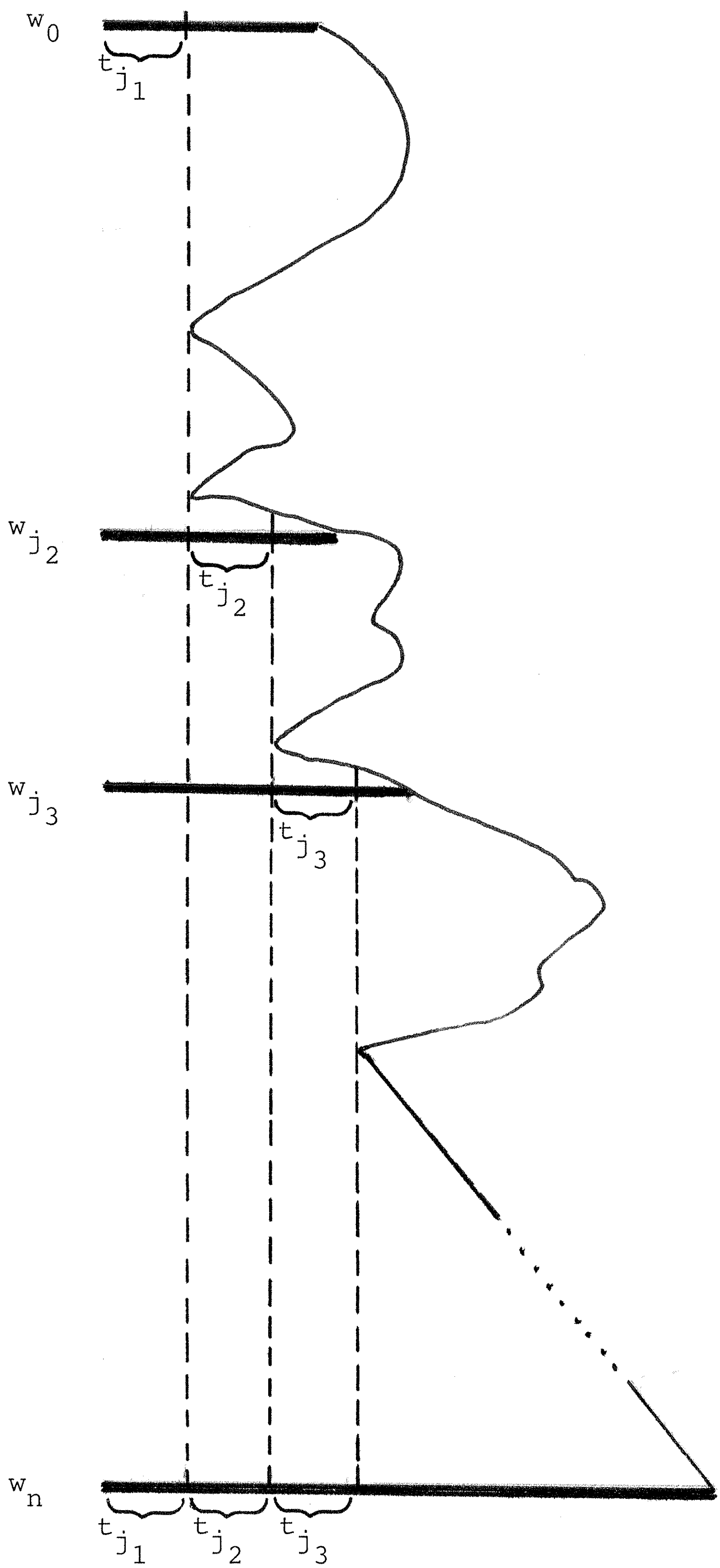

Figure 3

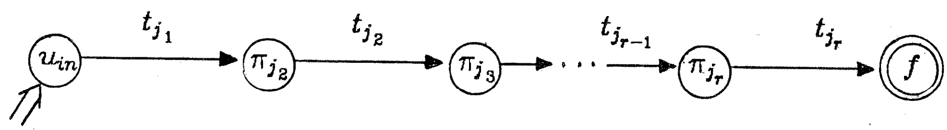

Figure 4.

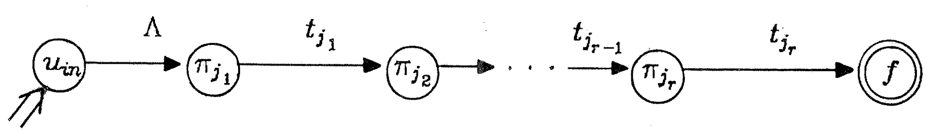

Figure 5.