CONDITIONS ENFORCING REGULARITY OF CONTEXT-FREE LANGUAGES

by

A. Ehrenfeucht
G. Rozenberg
D.Haussler

CU-CS-223-82

A. Ehrenfeucht
D. Haussler
Dept. of Computer Science
Univ. of Colo., Boulder
Boulder, Colorado 80309

G. Rozenberg
Institute of Applied Mathematics
and Computer Science
University of Leiden
Wassenaarsweg 80
2333 AL Leiden
The Netherlands

All correspondence to
G. Rozenberg.

CONDITIONS ENFORCING REGULARITY OF CONTEXT-FREE LANGUAGES

by

A. Ehrenfeucht
D. Haussler
Dept. of Computer Science
University of Colorado at Boulder
Boulder, Colorado 80309
U.S.A.

and

G. Rozenberg
Institute of Applied Mathematics
and Computer Science
University of Leiden
Wassenaarseweg 80
2333 AL Leiden
The Netherlands

The class of context-free languages (L_{CF}) and the class of regular languages (L_{REG}), where $L_{REG} \subsetneq L_{CF}$, are important classes of languages within formal language theory (see, e.g., [H] and [S]). In order to understand the relationship between "context-freeness" and "regularity" one can proceed in (at least) two different ways:
(1). Investigate conditions under which a context-free grammar will generate a regular language; several restrictions of this kind are known, the self-embedding property is a classical example of such a condition (see, e.g., [H] and [S]).
(2). Investigate conditions which imposed on (the interrelationship of words in) a context-free language will guarantee that the generated language is regular. Several conditions of this kind are known (see, e.g., [ABBL] and [ABBN]).

This paper presents several results concerning the second line of research discussed above.

1. STRONG ITERATIVE PAIRS.

A fundamental property of context-free languages is the celebrated pumping property (see, e.g., [H] and [S]). Based on it the notion of an iterative pair was introduced in [B] (see also [ABBL]). If K is a language, $K \subseteq \Sigma^*$ then $p = (x,y,z,u,t)$ with $x,y,z,u,t \in \Sigma^*$ is an iterative pair in K if, for every $n \geq 1$, $xy^nzu^nt \in K$ where yu is a nonempty word. Such a synchronized pumping of subwords (y and u) in a word (xyzut) of K gives one a possibility (using one iterative pair only) to generate context-free but not regular languages (e.g., $\{a^nb^n : n \geq 1\}$). However, if one desynchronizes such a pumping, that is one requires that for all $r,s \geq 0$, $xy^rru^st \in K$ then an iterative pair yields a regular language. This observation leads one to a conjecture that if each iterative pair $p = (x,y,z,u,t)$ of a context-free language K is very degenerate (that is, for all $r,s \geq 0$, $xy^rzu^st \in K$) then K must be regular. This conjecture was shown in [B] to be true. An iterative pair allows only "upward pumping" expressed by the fact that $n \geq 1$ and in this sense it does not fully forma-

lize the idea from the pumping lemma for context-free languages. There, also the "downward pumping" (i.e. n = 0) is allowed; it is well-known that this downward pumping is a very essential part of the pumping property for context-free languages.

If in the definition of an iterative pair we require "$n \geq 0$" rather than "$n \geq 1$" then we get a strong iterative pair. Then the "full version" of the conjecture mentioned above is:

Conjecture 1. If each strong iterative pair of a context-free language K is very degenerate then K is regular. □

We prove the following result.

Theorem 1. Conjecture 1 holds. □

The above result solves a problem remaining open since [B] ([B1] and [ABBL]). Also, Theorem 1 generalizes the above mentioned result from [B] which can be obtained directly from our theorem.

2. COMMUTATIVE LINEAR LANGUAGES.

Let for a word w, c(w) denote the commutative image of w, i.e., the set of all words that can be obtained from w be permuting (occurrences of) letters in it. For a language K, its commutative image is defined by $c(K) = \bigcup_{w \in K} c(w)$. We say that a language K is commutative if K = c(K). Commutative languages form a very active research topic within formal language theory (see, e.g., [ABBL], [L1], [L2] and [SS]). In the literature there are several conjectures known which relate regularity and commutativeness of a formal language (see, e.g., [ABBL] and [L1]).

Linear languages form perhaps a closest natural extension of regular languages; the only difference being that in generating the former one can insert substrings inside strings already generated (rather than one the edge of strings only as happens in right-linear grammars). It seems quite feasible that requiring a linear language being commutative removes (the consequences of) the difference mentioned above. Hence the following was conjectured ([L1] and [L3]).

Conjecture 2. If a language K is commutative and linear then it is regular. □

We prove that the above conjecture is true; as a matter of fact we prove a more general result.

Let $\Sigma = \{a_1,\ldots,a_d\}$, $d \geq 1$, be an arbitrary but fixed alphabet. Let $\rho = v_0,v_1,\ldots,v_d$ be a sequence of vectors each of which has d components where every component is a nonnegative integer. We say that ρ is a base if and only if $v_i(j) = 0$ for all $i,j \geq 1$ such that $i \neq j$. The ρ-set, denoted $\theta(\rho)$, is defined by $\theta(\rho) = \{v \in \Psi(\Sigma^*) : v = v_0 + \ell_1 v_1 + \ell_2 v_2 + \ldots + \ell_d v_d$ for some nonnegative integers $\ell_1,\ldots,\ell_d\}$,
where for a language K, $\Psi(K)$ denotes the set of Parikh vectors of K.

Let $X \subset \Psi(\Sigma^*)$. We say that X is periodic if and only if there exists a base ρ

such that $X = \theta(\rho)$. A language $K \subseteq \Sigma^*$ is periodic if and only if K is commutative and $\Psi(K)$ is periodic; the base of $\Psi(K)$ is also called the base of K and denoted base(K).

Let K be a periodic language where base(K) = $v_0, v_1, \ldots, v_d$. The size of K, denoted size(K), is defined by $size(K) = \max_{1 \le i \le d} \{\max\{u_1(i), u_2(i)\}\}$ where $type(K) = (u_1, u_2)$.

We prove the following result.

Theorem 2. Let $K \subseteq \Sigma^*$. If there exists a positive integer q such that for each $w \in K$ there exists a periodic language $L_w \subseteq K$ where $w \in L_w$ and size $(L_w) \le q$ then K is a finite union of periodic languages. □

Using this result we prove

Theorem 3. A language K is a commutative linear language if and only if K is a finite union of periodic languages. □

Since it is easily seen that each periodic language is regular the above result yields.

Theorem 4. Conjecture 2 holds. □

3. INCLUDING SQUARES.

A very fundamental structure of a string (or a language) is a repetition of its substrings. For example, a string x is said to be a pure-square if $x = yy$ where y is a nonempty string, x is a square if x contains a pure square as a subword and x is square-free if it is not a square. Such structures were for the first time systematically investigated by Thue ([T]) and later on in very many papers concerning various branches of mathematics (see, e.g., [Be], [BEM], [S] and references therein). These structures turned out to be of fundamental importance in formal language theory (see, e.g., [ABBL], [B2], [S]). It was proved recently (see [ER] and [RW]) that the set of all squares (over an alphabet containing at least three letters) is not a context-free language. This result (and its proofs) support the rahter old and very powerful conjecture (see, e.g., [ABBL]).

Conjecture 3. If a context-free language $K \subseteq \Delta$ contains all squares over Δ^* then K is regular. □

The intuition behind this conjecture is that if a context-free grammar generates all squares over Δ then it generates "almost all words" over Δ. We are not able to either prove or disprove this conjecture, however, we can prove that a somewhat weaker form of this conjecture is false.

Theorem 5. There exists a context-free language $K \subseteq \{a,b\}^*$ such that K contains all pure squares over {a,b} and K is not regular. □

4. INSERTION SYSTEMS.

Insertion systems formalize a very special type of semi-Thue systems. An insertion system is a triple $G = (\Delta, I, w)$ where Δ is a finite nonempty alphabet. I is a finite nonempty subset of Δ^+ and $w \in \Delta^*$; I is called the insertion set of G and w is called the axiom of G. If $w = \Lambda$ then we say that G is pure. For $u \in \Delta^*$, $v \in \Delta^+$ we say that u directly derives v (in G) if $u = u_1u_2$ for some $u_1, u_2 \in \Delta$ and $v = u_1zu_2$ where $z \in I$; we write then $u \overset{}{\Rightarrow}_G v$. Then $\overset{*}{\Rightarrow}_G$ denotes the transitive and the reflexive closure of the $\overset{}{\Rightarrow}_G$ relation; if $u \overset{*}{\Rightarrow}_G v$ then we say that u derives v (in G). The language of G, denoted L(G), is defined by $L(G) = \{v \in \Delta^* : w \overset{*}{\Rightarrow}_G v$; it is referred to as an insertion language or a pure insertion language if G is pure.

The insertion languages form a very natural generalization of restricted Dyck languages. Clearly the class of insertion languages strictly contains the class of restricted Dyck languages and it is strictly contained in the class of context-free languages.

In order to establish conditions under which an insertion language becomes regular we have to prove two results first. These results are of independent interest: the first of them generalizes the celebrated theorem by Higman (see [Hi]) on ordering of words by the sparse subword relationship, the second one provides a new algebraic characterization of regular languages. In order to state those results we need some additional terminology.

Let us recall (see, e.g., [Hi] and [N]) that a relation that is reflexive and transitive is called a quasi-order (qo). If $\leq$ is a quasi-order defined on a set S, then $\leq$ is called a well-quasi-order (wqo) if and only if any of the following holds.
(1). $\leq$ is well founded on S, i.e., there exist no infinite strictly descending sequences of elements in S and each set of pairwise incomparable elements is finite.
(2). For each infinite sequence $\{x_i\}$ of elements in S there exist $i < j$ such that $x_i \leq x_j$.
(3). Each infinite sequence of elements in S contains an ascending infinite subsequence.

Given a finite nonempty set of words $I \subset \Delta^+$ we say that I is subword complete if and only if there exists a positive integer m such that for each word z in Δ^* longer than m there exist $u,v \in \Delta^*$ and $w \in I$ such that $z = uwv$.
Let I be a finite nonempty subset of Δ^+. For $x,y \in \Delta^*$ we write $x \leq_I y$ if $x \overset{*}{\Rightarrow}_G y$ where G is the insertion system (Δ, I, x).

Theorem 6. Let I be a finite nonempty subset of Δ^+. Then $\leq_I$ is a well-quasi-order if and only if I is subword complete. □

A quasi-order $\leq$ on Δ^* is called monotone if and only if for all $x_1, x_2, y_1, y_2 \in \Delta^*$ the following holds: if $x_1 \leq y_1$ and $x_2 \leq y_2$ then $x_1x_2 \leq y_1y_2$. A set $S \subseteq \Delta^*$ is upwards closed under $\leq$ if and only if whenever $x \in S$ and $x \leq y$ then $y \in S$.

Theorem 7. Let $K \subseteq \Delta^*$. K is regular if and only if there exists a monotone wqo $\leq$

on Δ^* such that K is upwards closed under $\leq$. □

Using the above two results we can provide the following characterization of regular insertion languages.

Theorem 8. Let K be the insertion language generated by an insertion system $G = (\Delta, I, w)$. Then K is regular if and only if I is subword complete. □

ACKNOWLEDGEMENTS.

The authors gratefully acknowledge the support of NSF grant MSC 79-03838.

REFERENCES

[ABBL] J.M. Autebert, J. Beauquier, L. Boasson and M. Latteux, Very small families of algebraic nonrational languages, in R. Book (ed.), Formal language theory; perspectives and open problems, 1980, Academic Press, London, New York, 89-108.

[ABBN] J.M. Autebert, J. Beauquier, L. Boasson and M. Nivat, Quelques problèmes ouverts en théorie des languages algébriques, 1979, RAIRO Informatique Theorique, v. 13, 363-379.

[BEM] D.R. Bean, A. Ehrenfeucht and G.F. McNulty, Avoidable patterns in strings of symbols, 1979, Pacific Journal of Mathematics, v. 85, n.2, 261-293.

[Be] J. Berstel, Sur les mots sans carré definis par un morphisme, 1979, Springer Lecture Notes in Computer Science, v. 71, 16-25.

[B] L. Boasson, Un critére de rationnalité des languages algébriques, in M. Nivat (ed.), Automata, Languages and Programming, 1973, North-Holland, Amsterdam, 359-365.

[B1] L. Boasson, private communication.

[ER] A. Ehrenfeucht and G. Rozenberg, On the separating power of EOL systems, RAIRO Informatique Theorique, to appear.

[H] M. Harrison, Introduction to formal language theory, 1978, Addison-Wesley, Reading, Massachusetts.

[Hi] G.H. Higman, Ordering by divisibility in abstract algebras, 1952, Proc. London Math. Society, v.3, 326-336.

[L1] M. Latteux, Ph.D. thesis, 1979, University of Lille.

[L2] M. Latteux, Cônes rationnels commutatifs, 1979, Journal of Computer and Systems Science, v. 18, 307-333.

[L3] M. Latteux, private communication.

[NW] C.St.J.A. Nash-Williams, A survey of the theory of well-quasi-ordered sets, in Combinatorial Structures and Their Applications, 1970, Gordon and Breach, New York, London, 293-300.

[RW] R. Ross and K. Winklman, Repetitive strings are not context-free, RAIRO Informatique Theorique, to appear.

[S] A. Salomaa, Jewels of formal language theory, 1981, Computer Science Press, Rockville, Maryland.

[T] A. Thue, Über unendliche Zeichenreihen, 1906, Norske Vid. Selsk.Skr., I Mat. Nat. Kl., Christiania, v. 7, 1-22.