Theoretical Computer Science 27 (1983) 311-332
North-Holland

ON REGULARITY OF CONTEXT-FREE LANGUAGES*

A. EHRENFEUCHT

Department of Computer Science, University of Colorado, Boulder, CO 80309, U.S.A.

D. HAUSSLER

Department of Mathematics and Computer Science, University of Denver, 2360 S. Gaylord, Denver, CO 80208, U.S.A.

G. ROZENBERG

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

Abstract. This paper considers conditions under which a context-free language is regular and conditions which imposed on (productions of) a rewriting system generating a context-free language will guarantee that the generated language is regular. In particular:

(1) necessary and sufficient conditions on productions of a unitary grammar are given that guarantee the generated language to be regular (a unitary grammar is a semi-Thue system in which the left-hand of each production is the empty word), and

(2) it is proved that commutativity of a linear language implies its regularity.

To obtain the former result, we give a generalization of the Myhill-Nerode characterization of the regular languages in terms of well-quasi orders, along with a generalization of Higman's well-quasi order result concerning the subsequence embedding relation on Σ^*. In obtaining the latter results, we introduce the class of periodic languages, and demonstrate how they can be used to characterize the commutative regular languages. Here we also utilize the theory of well-quasi orders.

1. Introduction

The most extensively investigated language classes within formal language theory are undoubtedly the class of regular languages ($L(REG)$) and the more general class of context-free languages ($L(CF)$) (see, e.g., [14] and [24]). In order to understand the strict inclusion between these language classes (that is, to understand the difference between 'context-freeness' and 'regularity') one can proceed in (at least) two different ways:

(1) Investigate conditions under which a context-free *grammar*, or any other type of a grammar generating context-free languages, generates a regular language: 'right-linearity' and 'non-self-embedding' of context-free grammars are classic examples of conditions of this type (see, e.g., [14] and [24]).

* This research was supported by NSF Grant MSC 79-03838.

(2) Investigate conditions which imposed on (the interrelationship of words in) a context-free *language* will guarantee that the language is regular.

Several conditions of this latter kind are known (see, e.g., [2] and [3]), although it seems that this line of research has yielded fewer results than the former.

In this paper we present results of both types. That is, we present (in Section 4) a condition which imposed on certain types of grammars will imply that the generated languages are regular and we present (in Section 6) a condition which imposed on the interrelationship of words in a special type of a context-free language will guarantee that the language is regular. Underlying both of these results are specific instances (given in Sections 3 and 5) of the application of the theory of *well-quasi orders* (see, e.g., [17]) to the study of regularity in language theory. This unifying technical theme is discussed in more detail in Section 7.

At this point, let us discuss the precise nature of our results in more detail, beginning with those of the first type.

In one of the earliest papers in the area of formal language theory, Thue [25] defines and investigates the power of a simple class of rewriting systems (grammars) now known as *Thue systems*. In a Thue system, the rewriting operation is always taken to be a symmetric operation, that is, if it is specified that a word u may be rewritten by a word v, then it is understood that v may be rewritten by u as well. Later, Post [23] investigated a more general type of rewriting system known as a *semi-Thue system*.

Formally, we have the following.

Definition. A *semi-Thue system* (over the alphabet Σ) is a finite set of pairs of words $T=\{\langle u_1, v_1\rangle, \ldots; \langle u_k, v_k\rangle\}$ where $u_i, v_i \in \Sigma^*$ for $1 \leq i \leq k$. For $x, y \in \Sigma^*$ we say that x *directly derives* y (*in* T) if $x=x_1ux_2$ and $y=x_1vx_2$ where $x_1, x_2 \in \Sigma^*$ and $\langle u, v\rangle \in T$. In this case we write $x \Rightarrow_T y$. $\Rightarrow_T^*$ denotes the reflexive, transitive closure of the relation $\Rightarrow_T$; if $x \Rightarrow_T^* y$, then we say that x *derives* y (*in* T).

From our perspective, a Thue system is a special type of semi-Thue system in which it is required that $\langle v, u\rangle \in T$ whenever $\langle u, v\rangle \in T$. When the condition of symmetry is imposed, the relation $\Rightarrow_T^*$ becomes an equivalence relation on the monoid Σ^*. In the general case, however, we can only say that this derivation relation is reflexive and transitive. A relation of this type is called a *quasi order*. It is clear that the notion of a quasi order generalizes the notion of a partial order (by not demanding anti-symmetry) as well as the notion of an equivalence relation (by not demanding symmetry). A quasi order does not necessarily partition Σ^* into disjoint classes, but it does induce a structure on Σ^* in the following sense.

Definition. For any quasi order $\leq$ on a set A and any subset B of A, the *upward closure* of B (with respect to $\leq$) is given by $cl_{\leq}(B)=\{x \in A: y \leq x \text{ for some } y \in B\}$. B is $\leq$*-closed* (or simply *closed*) if $cl_{\leq}(B)=B$ (i.e., if $x \in B$ and $x \leq y$ implies that $y \in B$).

It is clear that the notion of a $\leqslant$-closed set generalizes the notion of an equivalence class in that for any equivalence relation $\equiv$ on a set A, the equivalence class of an element x of A is $cl_{\equiv}(x)$.

We will be primarily concerned with a special kind of semi-Thue system which induces a partial order on the monoid Σ^*. Here we borrow some of the terminology from the theory of (full) Thue systems (see e.g. [6]).

Definition. Given a finite set $I \subseteq \Sigma^+$, the *semi-Thue system induced by I* is defined by $S(I) = \{\langle\lambda, w\rangle : w \in I\}$, where λ is the null word. A semi-Thue system T is *unitary* if $T = S(I)$ for some finite $I \subseteq \Sigma^+$.

For brevity, we will let $\leqslant_I$ denote the quasi order $\Rightarrow^*_{S(I)}$.

We can think about the derivation relation $\leqslant_I$ in terms of the unrestricted, repeated insertion of words from the set I. Thus it is apparent that for $x, y \in \Sigma^*$ and $I \subseteq \Sigma^+$, $x \leqslant_I y$ if and only if $x = a_1 \cdots a_k$ and $y = u_1 a_1 u_2 a_2 \cdots u_k a_k u_{k+1}$ for some $a_1, \ldots, a_k \in \Sigma$ and $u_1, \ldots, u_{k+1} \in cl_{\leqslant_I}(\lambda)$. In the special case that $I = \Sigma$, $cl_{\leqslant_I}(\lambda) = \Sigma^*$ and this relation reduces to the subsequence embedding relation on Σ^*, studied in [16] and [13]. On the other hand, if we let $D_1 = \{a\bar{a} : a \in \Sigma\}$ and $D_2 = \{a\bar{a}, \bar{a}a : a \in \Sigma\}$ (where $\bar{\Sigma}$ is a 'shadow' alphabet in one-to-one correspondence with Σ), then $cl_{\leqslant_{D_1}}(\lambda)$ is the restricted Dyck language with parenthesis of type $a, \bar{a}$ (see, e.g., [14]) and $cl_{\leqslant_{D_2}}(\lambda)$ is the full Dyck language over the alphabet Σ, i.e., the set of all words which represent λ in the free group generated by Σ, where $\bar{a}$ represents a^{-1} (see e.g. [1]). Thus the class of languages of the form $cl_{\leqslant_I}(w)$ for a finite set $I \subseteq \Sigma^+$ and a word $w \in \Sigma^*$ in some sense constitutes a class of 'generalized Dyck languages'. Cochet and Nivat [7] investigate a similar class of generalized Dyck languages, which they define as the set of all equivalence classes generated by unitary (full) Thue systems which are 'perfect'. ('Perfect' Thue systems are also called (strict) *Church-Rosser* Thue systems, e.g., in [6]). Due to the special nature of perfect Thue systems, any equivalence class in a unitary system of this type can be represented as $cl_{\leqslant_I}(w)$ for some word w, where $\leqslant_I$ is the derivation relation of the corresponding unitary semi-Thue system. Thus our notion of a 'generalized Dyck language' includes those languages of Cochet and Nivat.

Let us formalize these concepts by defining a class of grammars associated with the derivation relation $\leqslant_I$.

Definition. A *unitary grammar* is a triple $G = \langle\Sigma, I, w\rangle$ where Σ is a finite, nonempty alphabet, $I \subseteq \Sigma^+$ is a finite, nonempty set and $w \in \Sigma^*$. I is called the *insertion set* of G and w is called the *axiom* of G. The *language of G*, denoted $L(G)$, is $cl_{\leqslant_I}(w)$. If $w = \lambda$, then G is called a *pure* unitary grammar. A language of the form $L(G)$ for some unitary grammar G is called a *unitary language* or a *pure unitary language* if G is pure.

It is apparent that unitary languages are context-free; the construction of a context-free grammar generating the language of a given unitary grammar is obvious. Thus the classes of unitary languages and pure unitary languages form natural generalizations of the class of Dyck languages within the class of context-free languages. These classes are, of course, properly contained within the class of context-free languages, because the unitary languages are always infinite. Furthermore, using the Chomsky–Schutzenberger representation theorem [5], any context-free language is the homomorphic image of the intersection of a unitary language with a regular set. Since the regular sets are closed under intersection and homomorphism, this indicates that at least some unitary languages capture part of the essential 'non-regular' aspects of the context-free languages. On the other hand, it is clear that if the insertion set I is 'too dense', for example, if $\Sigma \subseteq I$, then the unitary languages generated by I will themselves be regular sets. We will investigate the conditions under which a unitary grammar generates a regular language. Our results may be briefly outlined as follows.

In Section 3 we give a general condition on a semi-Thue system T which enforces the regularity of any upward closed set with respect to $\Rightarrow_T^*$. This is accomplished by generalizing the Myhill–Nerode characterization of the regular sets, given in terms of equivalence relations of finite index, to the more general class of well-quasi orders. Relevant definitions are given in the first part of this section. Following this, in Section 4 we explore the specific conditions which ensure the regularity of unitary languages. By generalizing the theorem of Higman [16] which shows that the subsequence embedding relation $\leqslant_\Sigma$ is a well-quasi order on Σ^*, we are able to give necessary and sufficient conditions under which a unitary grammar generates a regular set.

In Sections 5 and 6 we turn to results of the second type discussed above, that is, we present results concerning conditions which imposed on a context-free *language* will guarantee that the language is regular.

In an effort to learn more about such conditions one may investigate subclasses of $L(CF)$ which are 'as small as possible' (and still contain $L(REG)$). One such class is the class of linear languages $L(LIN)$. A linear grammar differs from a right-linear grammar (which always generates a regular language) only by the fact that the unique nonterminal in a sentenial form may generate terminal symbols both to the right and to the left of itself. Hence it looks quite plausible that requiring commutativity of a linear language (that is, requiring that for every word each permutation of occurrences of letters in it will result in a word also in the language) will force the language to be regular. This conjecture was formulated in [18] where various properties of commutative context-free languages are considered. In Section 6 we demonstrate that this conjecture holds. Actually, our result is more general in that it relates commutative linear languages to periodic languages, which are introduced and investigated in Section 5. In this investigation well-quasi orders turn out to be an important technical tool as well.

2. Preliminaries

We assume the reader to be familiar with the basic theory of context-free languages; in particular with the basic theory of regular and linear languages (see, e.g., [24]). We use mostly standard language theoretic terminology and notation. Perhaps the following points require an additional explanation.

We use $\mathbb{N}$ to denote the set of nonnegative integers and $\mathbb{N}^+$ to denote the set of positive integers. For $n \in \mathbb{N}^+$, $\mathbb{N}^n$ denotes the n-fold cartesian product of $\mathbb{N}$. If $v \in \mathbb{N}^n$, then, for $1 \leq i \leq n$, $v(i)$ denotes the ith component of v. If $v_1, v_2 \in \mathbb{N}^n$, then $v_1 \leq v_2$ if and only if $v_1(i) \leq v_2(i)$ for each $1 \leq i \leq n$.

For a finite set Z, $\#Z$ denotes its cardinality. For sets Z_1, Z_2, $Z_1 - Z_2$ denotes the set-theoretic difference of Z_1 and Z_2.

For a word w, $alph(w)$ denotes the set of all letters that occur in w and $|w|$ denotes the length of w. For a letter a and a word w, $\#_a(w)$ denotes the number of occurrences of a in w. For $\Sigma = \{a_1, \ldots, a_d\}$, $\Psi : \Sigma^* \to \mathbb{N}^d$ is the mapping defined by $\Psi(w) = (\#_{a_1}(w), \ldots, \#_{a_d}(w))$ for all $w \in \Sigma^*$; Ψ is referred to as the *Parikh mapping* and $\Psi(w)$ as the *Parikh vector* of w. For $K \subseteq \Sigma^*$, $\Psi(K) = \bigcup_{w \in K} \Psi(w)$.

3. A generalization of the Myhill–Nerode characterization of the regular sets

The importance of certain equivalence relations of finite index in the theory of regular sets was first observed by Myhill [20] and Nerode [22].

Definition. A binary relation R on a finitely generated free monoid Σ^* is *monotone* if and only if $x_1 R y_1$ and $x_2 R y_2$ implies that $x_1 x_2 R y_1 y_2$ for all $x_1, x_2, y_1, y_2 \in \Sigma^*$. A monotone equivalence relation is called a *congruence*.

Proposition 3.1 (Myhill-Nerode characterization). *For any set $S \subseteq \Sigma^*$, S is regular if and only if S is a union of equivalence classes under some congruence on Σ^* of finite index.*

In the terminology introduced in Section 1, this result states that a set S is regular if and only if S is $\equiv$-closed under some monotone equivalence relation $\equiv$ of finite index. Our goal is to investigate the regularity of sets which are closed under the derivation relations of semi-Thue systems. These are monotone relations, but in general they are only quasi orders, and not full equivalence relations. Thus, to investigate the regularity of sets generated by such relations, we would like to find some generalization of the equivalence relations of finite index in the class of quasi orders. This is the class of *well-quasi orders* (see, e.g., [17]). In his seminal paper, Higman [16] gives the following definitions and proves them to be equivalent.

Definition. Given a set A and a quasi order $\leqslant$ on A, then $\leqslant$ is a *well-quasi order* on A if and only if any of the following hold:

(i) $\leqslant$ is well founded on A (i.e., there exist no infinite descending sequences $a_1 \geqslant a_2 \geqslant \cdots$ such that $a_i \not\leqslant a_{i+1}$ for any i) and each set of pairwise incomparable elements in A is finite,

(ii) for each infinite sequence $\{x_i\}$ of elements in A, there exist $i<j$ such that $x_i \leqslant x_j$,

(iii) each infinite sequence of elements in A contains an infinite ascending subsequence,

(iv) A has the 'finite basis property', i.e., for each set $C \subseteq A$ there exists a finite $B_C \subseteq C$ such that for every $c \in C$ there exists a $b \in B_C$ such that $b \leqslant c$, and

(v) every sequence of $\leqslant$-closed subsets of A which is strictly ascending under inclusion is finite.

From the above definition (i) it is obvious that the class of equivalence relations of finite index is exactly the class of symmetric well-quasi orders. Thus the class of congruences of finite index is exactly the class of symmetric monotone well-quasi orders. We generalize the Myhill–Nerode characterization to show that a set is regular if any only if it is upward closed with respect to some arbitrary monotone well-quasi order. Our proof mirrors the basic technique used in [8] to show that sets closed under the subsequence embedding relation are regular. It is derived from the following alternate characterization of the regular sets, usually attributed to Nerode (see, e.g., [14]).

Definition. Given $L \subseteq \Sigma^*$ and $w \in \Sigma^*$, $F_L(w) = \{x : wx \in L\}$. The equivalence relation $\equiv_L$ on Σ^* is defined by $x \equiv_L y$ if and only if $F_L(x) = F_L(y)$. The equivalence classes induced by $\equiv_L$ are called the *right invariant* equivalence classes of L.

Proposition 3.2. *For any $L \subseteq \Sigma^*$, L is regular if and only if $\equiv_L$ partitions Σ^* into a finite number of distinct equivalence classes.*

Theorem 3.3 (generalized Myhill–Nerode characterization). *For any $S \subseteq \Sigma^*$, S is regular if and only if S is $\leqslant$-closed under some monotone well-quasi order $\leqslant$ on Σ^*.*

Proof. Since the 'only if' part follows from Proposition 3.1, it suffices to show that for any monotone well-quasi order $\leqslant$ on Σ^* each $\leqslant$-closed set S in Σ^* is regular. Let us assume to the contrary that we are given a monotone well-quasi order $\leqslant$ on Σ^* and a $\leqslant$-closed set $S \subseteq \Sigma^*$ which is not regular. Since S is not regular the number of distinct right invariant equivalence classes under $\equiv_S$ must be infinite by Proposition 3.2. Hence we can find an infinite sequence $\{w_i\}$ of words in Σ^* such that $w_i \not\equiv_S w_j$ for $i \neq j$. Since $\leqslant$ is a well-quasi order, there exists an infinite subsequence of $\{w_i\}$ which is ascending with respect to $\leqslant$, using the above definition (iii). Hence, we may assume that $\{w_i\}$ itself is chosen as an ascending sequence.

Since $\leqslant$ is monotone and $\{w_i\}$ is ascending, for any $x \in \Sigma^*$ and $i<j$, $w_ix \leqslant w_jx$. Hence since S is $\leqslant$-closed, $w_ix \in S$ implies that $w_jx \in S$. Thus the sequence $\{F_S(w_i)\}$ is ascending with respect to inclusion. Further, since $w_i \not\equiv_S w_j$ for $i \neq j$, $\{F_S(w_i)\}$ is strictly ascending. Now, by the same reasoning as above, for any i and any $x \leqslant y$, if $w_ix \in S$, then $w_iy \in S$. Hence, each $F_S(w_i)$ is $\leqslant$-closed. Thus $\{F_S(w_i)\}$ forms an infinite strictly ascending sequence of $\leqslant$-closed sets, contradicting the fact that $\leqslant$ is a well-quasi order, using the above definition (v). We conclude that every $\leqslant$-closed set S is regular. □

Corollary 3.4. *For any semi-Thue system T over an alphabet Σ, if $\Rightarrow_T^*$ is a well-quasi order on Σ^*, then $cl_{\Rightarrow_T^*}(S)$ is regular for every $S \subseteq \Sigma^*$.*

Proof. Since $\Rightarrow_T^*$ is monotone for any semi-Thue system T, this follows directly from the above theorem. □

4. A characterization of regularity in unitary grammars

It is clear from Corollary 3.4 that a unitary grammar $G = \langle \Sigma, I, w \rangle$ will generate a regular set whenever $\leqslant_I$ is a well-quasi order on Σ^*. Under what conditions does this occur? It has been known for some time that in the special case that $I = \Sigma$, $\leqslant_I$ defines a well-quasi order on Σ^*. This result was given in [16]. Conway [8] gives a very elegant proof of the result using a technique originally due to Nash-Williams [21].

To be definite, let us suppose that $\Sigma = \{a, b\}$. It is not hard to extend Conway's method to show that if $I = \{aa, ab, ba, bb\}$, then $\leqslant_I$ is a well-quasi order on Σ^*. In fact we can show that this is the case when $I = \Sigma^k$ for any $k \geqslant 1$. The case $I = \{aa, ab, bb\}$ is somewhat more difficult to handle. (It also generates a well-quasi order.) On the other hand, it is clear that if $I = \{aa, bb\}$, then $\leqslant_I$ will not be a well-quasi order on Σ^* because all of the words in $(ab)^+$ are pairwise incomparable with respect to $\leqslant_I$ (see part (i) of the definition of a well-quasi order in Section 3). This is because none of these words contains a subword which is in I, and thus no word in this set can be derived from any other word by any nonempty sequence of insertions from I. Let us generalize this example as follows.

Definition. A set $I \subseteq \Sigma^*$ is *subword avoidable* in Σ^* if and only if there exists an infinite subset S of Σ^* such that no $w \in S$ has a subword which is in I. Otherwise, I is *subword unavoidable* in Σ^*. If I is subword unavoidable in Σ^*, the smallest $k_0 \in \mathbb{N}$ such that all words longer than k_0 have a subword in I is called the *subword avoidance bound* for I.

By the above reasoning it is obvious that for $\leqslant_I$ to be a well-quasi order on Σ^*, it is necessary that I be subword unavoidable in Σ^*, because otherwise we can find an infinite subset of Σ^* whose elements are all pairwise incomparable. We demonstrate that this condition is in fact both necessary and sufficient. We begin by giving a few basic facts about well-quasi orders. Our first two propositions are immediate consequences of Higman's definitions, given in the previous section.

Proposition 4.1. *If $\leqslant_1$ is a well-quasi order on the set S and $\leqslant_2$ is an extension of $\leqslant_1$ which is also a quasi order, then $\leqslant_2$ is a well-quasi order on S.*

Definition. Given sets S_1 and S_2 and relations R_1 and R_2 on S_1 and S_2 respectively, the relation $R_1 \times R_2$ on $S_1 \times S_2$ is defined by $\langle a, b\rangle R_1 \times R_2 \langle c, d\rangle$ if and only if $a R_1 c$ and $b R_2 d$.

Proposition 4.2. *Given sets S_1, S_2 and well-quasi orders $\leqslant_1$ and $\leqslant_2$ on S_1 and S_2 respectively, the transitive closure of $\leqslant_1 \cup \leqslant_2$ is a well-quasi order on $S_1 \cup S_2$ and $\leqslant_1 \times \leqslant_2$ is a well-quasi order on $S_1 \times S_2$.*

One of the earliest results of the theory of well-quasi orders is the following, apparently discovered independently by Higman, Neumann and Erdös and Rado around 1950 (see note at the end of [11]).

Definition. For any set S, $S^{<\omega}$ is the set of finite sequences of elements of S. Given a set S and a quasi order $\leqslant$ on S, the ordering $\leqslant^F$ on $S^{<\omega}$ is defined by $\langle s_1, \ldots, s_k\rangle \leqslant^F \langle t_1, \ldots, t_l\rangle$ if and only if there exists a subsequence $\langle t_{i_1}, \ldots, t_{i_k}\rangle$ of $\langle t_1, \ldots, t_l\rangle$ such that $s_j \leqslant t_{i_j}$ for $1 \leqslant j \leqslant k$.

Proposition 4.3. *If $\leqslant$ is a well-quasi order on S, then $\leqslant^F$ is a well-quasi order on $S^{<\omega}$.*

Proof. See [19] for a very short proof of this result. □

From these fundamental results concerning well-quasi orders in general, we can derive a few basic results concerning the special case of monotone well-quasi orders on subsets of Σ^*.

Lemma 4.4. *Given $S_1, S_2 \subseteq \Sigma^*$ and a monotone quasi order $\leqslant$ on Σ^*, if $\leqslant$ is a well-quasi order on S_1 and on S_2, then $\leqslant$ is a well-quasi order on S_1S_2.*

Proof. Let $\{x_iy_i\}$ be an infinite sequence of words in S_1S_2 where, for all i, $x_i \in S_1$ and $y_i \in S_2$. By Proposition 4.2, $\leqslant \times \leqslant$ is a well-quasi order on $S_1 \times S_2$. Thus we can find i, j such that $i < j$ and $(x_i, y_i) \leqslant \times \leqslant (x_j, y_j)$, i.e., $x_i \leqslant x_j$ and $y_i \leqslant y_j$. Since $\leqslant$ is monotone, this implies that $x_iy_i \leqslant x_jy_j$. Thus $\leqslant$ is a well-quasi order on S_1S_2 using part (ii) of the definition of a well-quasi order from Section 3. □

Lemma 4.5. *Given $S \subseteq \Sigma^*$ and a monotone quasi order $\leqslant$ on Σ^* where $\lambda \leqslant x$ for all $x \in S$, if $\leqslant$ is a well-quasi order on S, then $\leqslant$ is a well-quasi order on S^*.*

Proof. Let $\{u_{i,1} \cdots u_{i,k_i}\}$ be an infinite sequence of words in S^* where $u_{i,n} \in S$ for all i and all n, $1 \leq n \leq k_i$. Since $\leq$ is a well-quasi order on S, $\leq^E$ is a well-quasi order on $S^{<w}$ by Proposition 4.3. Thus we can find i, j such that $i < j$ and $\langle u_{i,1}, \ldots, u_{i,k_i} \rangle \leq^E \langle u_{j,1}, \ldots, u_{j,k_j} \rangle$. Hence there exists a subsequence $\langle u_{j,l_1}, \ldots, u_{j,l_{k_i}} \rangle$ of $\langle u_{j,1}, \ldots, u_{j,k_j} \rangle$ such that $u_{i,n} \leq u_{j,l_n}$ for $1 \leq n \leq k_i$. Since $\lambda \leq x$ for all $x \in S$, this implies that $u_{i,1} \cdots u_{i,k_i} \leq u_{j,1} \cdots u_{j,k_j}$ by monotonicity. Hence $\leq$ is a well-quasi order on S^*. □

Note that since the subsequence embedding relation $\leq_\Sigma$ is monotone and, for all $a \in \Sigma$, $\lambda \leq_\Sigma a$, the Higman result cited above can easily be derived from the above lemma. The heart of the argument used in the general case follows in the next two lemmas.

Definition. For each finite $I \subseteq \Sigma^*$ let

$$I_0 = I^* \quad \text{and} \quad I_{n+1} = \left[\bigcup_{a_1 \cdots a_k \in I - \{\lambda\}} I_n a_1 I_n a_2 \cdots I_n a_k I_n \right]^* .$$

Lemma 4.6. *For any finite $I \subseteq \Sigma^*$ and $n \geq 0$,*

(i) *if $uv \in I_n$ and $w \in I$, then $uwv \in I_{n+1}$,*

(ii) *if $uv \in I_n$, where $|u| \leq n$, and $w \in I$, then $uwv \in I_n$, and*

(iii) *$\leq_I$ is a well-quasi order on I_n.*

Proof. ad (i). This is obvious.

ad (ii). Here we use induction on n. If $n = 0$, we need only consider the case $u = \lambda$ and the statement follows from the fact that $I_0 = I^*$. Now let us assume that the statement holds for some $n \geq 0$. If $uv \in I_{n+1}$, then $uv = w_1 a_1 w_2 a_2 \cdots w_k a_k w_{k+1}$ where $w_i \in I_n$ for $1 \leq i \leq k+1$ and $a_1 \cdots a_k \in I^*$. Hence, for some i, $1 \leq i \leq k+1$, $u = w_1 a_1 \cdots w_{i-1} a_{i-1} w'_i$ and $v = w''_i a_i \cdots w_k a_k w_{k+1}$ where $w'_i, w''_i \in \Sigma^*$ and $w'_i w''_i = w_i$. For any $w \in I$, $w'_i w w''_i \in I_{n+1}$ by part (i). Thus if $i = 1$, then $uwv \in I_{n+1}$ because $a_1 w_2 \cdots a_k w_{k+1} \in I_{n+1}$ and I_{n+1} is closed under concatenation. On the other hand, it is apparent that if $i > 1$ and u has at most $n+1$ letters, w'_i has at most n letters. Thus by the inductive hypothesis, for any $w \in I$, $w'_i w w''_i \in I_n$. But this implies that $uwv \in I_n a_1 \cdots I_n a_k I_n$, and thus $uwv \in I_{n+1}$. Thus the statement holds for $n+1$ and the result follows by induction.

ad (iii). Again we use induction on n. We will need to induct on the stronger assertion that for every n, $\leq_I$ is a well-quasi order on I_n and $\lambda \leq_I w$ for every $w \in I_n$. First note that $\lambda \leq_I w$ for all $w \in I$ and, since I is a finite set, $\leq_I$ is a well-quasi order on I. Thus by Lemma 4.5, $\leq_I$ is a well-quasi order on I^*. Obviously $\lambda \leq_I w$ for every $w \in I^*$, thus the assertion holds for the case $n = 0$. Let us suppose this assertion holds for some $n \geq 0$. Using Lemma 4.4 we have that $\leq_I$ is a well-quasi order on $I_n a_1 \cdots I_n a_i I_n$ for any $a_1 \cdots a_i \in \Sigma^*$. Furthermore, if $a_1 \cdots a_i \in I \cup \{\lambda\}$, then, for any $w \in I_n a_1 \cdots I_n a_i I_n$, $\lambda \leq_I w$. Also, since I is finite, $\leq_I$ is a well-quasi order on $S = \bigcup_{a_1 \cdots a_i \in I \cup \{\lambda\}} I_n a_1 \cdots I_n a_i I_n$ using Proposition 4.2. Hence, by Lemma 4.5, $\leqslant_I$ is a well-quasi order on S^*. Furthermore, $\lambda \leqslant x$ for all $x \in S^*$. Since $S^* = I_{n+1}$, we have shown that the assertion holds for $n+1$. The result follows by induction. □

Definition. Given a finite set $I \subseteq \Sigma^+$, for each $n \in \mathbb{N}$, let $R(I_n) = \bigcup_{a_1, \ldots, a_k \in \Sigma, k \leqslant n} I_n a_1 I_n a_2 \cdots I_n a_k I_n$.

Lemma 4.7. *For any finite $I \subseteq \Sigma^+$,*

(i) *$\leqslant_I$ is a well-quasi order on $R(I_n)$ for all n,*

(ii) *$\{R(I_n)\}$ is an ascending sequence of sets such that $\Sigma^* = \bigcup_{n=1}^{\infty} R(I_n)$, and*

(iii) *if I is subword unavoidable in Σ^* and k_0 is the subword avoidance bound for I, then $\Sigma^* = R(I_{k_0})$.*

Proof. ad (i). This follows from Lemma 4.6(iii) (using Lemma 4.4 and Proposition 4.2 as above).

ad (ii). This is obvious.

ad (iii). Assume to the contrary that $\Sigma^* - R(I_{k_0}) \neq \emptyset$. Let x be among the shortest words in $\Sigma^* - R(I_{k_0})$. Since $R(I_{k_0})$ contains all words of length k_0 or less, x must be longer than k_0 letters. Since k_0 is the subword avoidance bound for I, we can find among the first $k_0 + 1$ letters of x a subword in I. Thus $x = uwv$ where $w \in I$ and u has k_0 or fewer letters. Since x was of minimal length, $uv \in R(I_{k_0})$. Hence $uv = w_1 a_1 \cdots w_k a_k w_{k+1}$ where $a_i \in \Sigma$ for $1 \leqslant i \leqslant k$ and $w_i \in I_{k_0}$ for $1 \leqslant i \leqslant k+1$. Find i such that $u = w_1 a_1 \cdots w_i'$, $v = w_i'' a_i \cdots w_k a_k w_{k+1}$ and $w_i' w_i'' = w_i$. Since $|u| \leqslant k_0$, $|w_i'| \leqslant k_0$. Thus, by Lemma 4.6(ii), $w_i' w w_i'' \in I_{k_0}$. Hence $x \in R(I_{k_0})$, contrary to hypothesis. □

Theorem 4.8 (generalized Higman theorem). *Given a finite set $I \subseteq \Sigma^+$, $\leqslant_I$ is a well-quasi order on Σ^* if and only if I is subword unavoidable in Σ^*.*

Proof. If I is not subword unavoidable in Σ^*, then we can find an infinite set of words in Σ^*, none of which contains a subword which is in I. It is obvious then that none of these words can be properly derived from any other word by insertions of words from I and thus all these words are pairwise incomparable with respect to $\leqslant_I$. It follows from definition (i) of Section 3 that $\leqslant_I$ is not a well-quasi order on Σ^*. On the other hand, if I is subword unavoidable in Σ^*, then I has some subword avoidance bound k_0 and thus, by Lemma 4.7(i) and (iii), $\leqslant_I$ is a well-quasi order on $R(I_{k_0})$ and $R(I_{k_0}) = \Sigma^*$. □

From the generalized Myhill–Nerode and Higman theorems, it is clear that the subword unavoidability of I is a sufficient condition to ensure that the language generated by any unitary grammar with insertion set I will be regular. Let us show that this is a necessary condition as well. We will use the following elementary fact concerning regular sets.

Definition. Given a language $L \subseteq \Sigma^*$, a *suffix bound* for L is a number $B \in \mathbb{N}$ such that for all $x \in \Sigma^*$, if $xy \in L$ for some $y \in \Sigma^*$, then $xy' \in L$ for some $y' \in \Sigma^*$ where $|y'| \leqslant B$.

Proposition 4.9. *For any $L \subseteq \Sigma^*$, if L is regular, then L has a suffix bound.*

Proof. From each right invariant equivalence class $[x]_{\equiv_L}$ such that $F_L(x) \neq \emptyset$, choose a word $w \in F_L(x)$. Let T be the set of words chosen. Let $B = \max_{w \in T}\{|w|\}$. Since T is finite by Proposition 3.2, B is finite. Obviously B has the required properties. □

Definition. Given a finite alphabet Δ, $I \subseteq_{\min} \Delta^*$ if and only if $I \subseteq \Delta^*$ and $I \not\subseteq \Sigma^*$ for any Σ properly contained in Δ. $I \subseteq_{\min} \Delta^+$ denotes the fact that $I \subseteq_{\min} \Delta^*$ and $\lambda \notin I$.

Definition. For any $L \subseteq \Delta^*$, L is prefix complete for $\Sigma \subseteq \Delta$ if and only if for every $w \in \Sigma^*$ there exist $x \in \Delta^*$ such that $wx \in L$.

Lemma 4.10. *For any finite $\Sigma \subseteq \Delta$, $I \subseteq_{\min} \Sigma^+$ and $w \in \Delta^*$, if $S = L(\langle \Delta, I, w \rangle)$ is regular, then S is prefix complete for Σ.*

Proof. Clearly, it suffices to show that for each $a \in \Sigma$, there exists an $x \in \Sigma^*$ such that $ax \in I$. Let us assume to the contrary that there exists an $a \in \Sigma$ such that $ax \notin I$ for any $x \in \Sigma^*$. Let $r_a(x) = \#_a(x)/|x|$ for any $x \in \Delta^+$. Let $M = \max_{v \in I}\{r_a(v)\}$. Since $I \subseteq_{\min} \Sigma^*$, $M > 0$. Using M we can obtain an upper bound on $r_a(x)$ for $x \in S$. In the worst case we have $w = a^k$ for some $k \geqslant 0$, which implies that $r_a(x) \leqslant (M(|x| - k) + k)/|x|$ for any $x \in S$. Since $(M(|x| - k) + k)/|x| \leqslant M + k/|x|$, this yields

$$r_a(x) \leqslant M + k/|x| \tag{1}$$

for any $x \in S$.

Since I is finite, it is apparent that $r_a(v_0) = M$ for some $v_0 \in I$. Find $b \in \Sigma - \{a\}$ and $z \in \Sigma^*$ such that $v_0 = bz$. Since $r_a(v_0) > 0$, we must have $z \neq \lambda$. Now let n and m be chosen such that $m > n + k$. Let $x = b^n z^m w$. Since $m > n + k$, $|bz|m > |z|m + n + k = |x|$. Hence $|bz|m/|x| > 1$, which implies that $M|bz|m/|x| > M$, and thus $r_a(x) = (M|bz|m + k)/|x| > M + k/|x|$. Thus $x \notin S$ by (1). On the other hand, it is obvious that $b^n z^n w \in S$ for all $n \in \mathbb{N}$. It follows that for any n, m such that $m > n + k$, $b^n \not\equiv_S b^m$ which implies that S is not regular by Proposition 3.2. The result follows. □

Definition. For any finite, nonempty $I \subseteq \Sigma^*$, let $l_I = \max_{v \in I}\{|v|\}$.

Lemma 4.11. *For any finite, nonempty $I \subseteq \Sigma^+$ and $w \in \Sigma^*$, if $uv \in S = L(\langle \Sigma, I, w \rangle)$ and $|u| > (l_I - 1)|v| + |w|$, then u has a subword in I.*

Proof. We use induction on $|v|$. If $|v| = 0$, then $u \in S$ and $|u| > |w|$, hence u has a subword in I. Now assume that the statement holds whenever $|v| \leqslant n$. Given $uv \in S$

such that $|u| > (l_I - 1)|v| + |w|$ and $|v| = n + 1$, find $xy \in S$ and $z \in I$ such that $xzy = uv$. (This is possible because $|uv| > |w|$.) If u is not a proper prefix of xz, then obviously u has a subword in I. If u is a proper prefix of xz but not of x, then $|y| \leq |v| - 1$ and $|u| \leq |x| + l_I - 1$. Hence $|x| + l_I - 1 > (l_I - 1)|v| + |w|$, which implies that $|x| > (l_I - 1)(|v| - 1) + |w|$, which implies that $|x| > (l_I - 1)|y| + |w|$. Thus since $|y| \leq n$, by the inductive hypothesis, x has a subword in I. Since x is a prefix of u, this implies that u has a subword in I. Finally, if u is a proper prefix of x, then find $x' \in \Sigma^+$ such that $x = ux'$. Then

$$|x'y| < |v|, \; xy = ux'y \in S \quad \text{and} \quad |u| > (l_I - 1)|v| + |w| \geq (l_I - 1)|x'y| + |w|.$$

Hence u has a subword in I by the inductive hypothesis. Thus the statement holds if $|v| = n + 1$. The result follows by induction. □

Theorem 4.12 (regularity characterization). *For any finite $\Sigma \subseteq \Delta$, $I \subseteq_{\min} \Sigma^+$ and $w \in \Delta^*$, $L(\langle \Delta, I, w \rangle)$ is regular if and only if I is subword unavoidable in Σ^*.*

Proof. Let $S = L(\langle \Delta, I, w \rangle)$. By Lemma 4.10 we know that if S is regular, then S is prefix complete for Σ. Furthermore, there also exists some suffix bound B for S by Proposition 4.9. Thus for any $x \in \Sigma^*$ there exists an $y \in \Delta^*$, $|y| \leq B$, such that $xy \in S$. Let $k_0 = B(l_I - 1) + |w|$. Then for any $x \in \Sigma^*$ such that $|x| > k_0$ there exists an $y \in \Delta^*$ such that $xy \in S$ and $|x| > (l_I - 1)|y| + |w|$. Hence, by Lemma 4.11, every such x has a subword in I. Thus I is subword unavoidable in Σ^* with subword avoidance bound less than or equal to k_0. This establishes the 'only if' part. In the other direction, it is clear that by the generalized Myhill–Nerode and Higman theorems, $R = cl_{\leq_I}(\lambda)$ is a regular set if I is subword unavoidable in Σ^*. Thus since $S = Ra_1Ra_2 \cdots Ra_kR$, where $a_1, \ldots, a_k \in \Delta$ and $w = a_1 \cdots a_k$, S is regular. □

It should be noted that the above theorem gives an effective test for the regularity of unitary languages. This follows from the fact that it is easily decidable whether or not a finite set is subword unavoidable. In fact, we have the following stronger result.

Theorem 4.13. *For any regular set $R \subseteq \Sigma^+$, it is decidable whether or not R is subword unavoidable in Σ^*.*

Proof. For any set $R \subseteq \Sigma^+$, the set of all words which do not have subwords in R is $\Sigma^* - \Sigma^* R \Sigma^*$. R is subword unavoidable if and only if this set is finite. This is easily decidable for regular R (see, e.g., [14]). □

5. Well-quasi orders and periodic languages

In this section we turn to the investigation of conditions which imposed on a context-free *language* will imply its regularity. The main result of this investigation, proved in the next section, is that each commutative linear language is regular.

In order to obtain this result, we need to develop a few fundamental tools for the investigation of regularity in commutative languages. These tools arise from the investigation of the properties of the class of *periodic languages*, which we will define shortly. Again, it is the theory of well-quasi orders that is applied to derive the fundamental properties of periodic languages used in our regularity results.

To simplify the notation, here and in the following section, we adopt the following convention:

> *all languages we consider will be over an arbitrary but fixed alphabet* $\Sigma = \{a_1, \ldots, a_d\}$ *where* $d \geq 2$.

We begin by formally defining the commutative languages.

Definition. (i). Let $w \in \Sigma^*$. The *commutative closure of* w, denoted $com(w)$, is defined by $com(w) = \{x \in \Sigma^*\colon \Psi(x) = \Psi(w)\}$.

(ii) A language K is *commutative* if $com(w) \subseteq K$ for each $w \in K$.

(iii) Let $X \subseteq \Psi(\Sigma^*)$. The *language of* X, denoted $L(X)$, is defined by $L(X) = \{w \in \Sigma^*\colon \Psi(w) \in X\}$.

The following result is a direct consequence of the above definition.

Lemma 5.1. (i). *Let* K_1, K_2 *be commutative languages,* $K_1 \subseteq K_2$ *if and only if* $\Psi(K_1) \subseteq \Psi(K_2)$.

(ii). *Let* $X \subseteq \Psi(\Sigma^*)$. *Then* $L(X)$ *is uniquely defined.*

We turn now to the definition of the periodic languages.

Definition. Let $\rho = v_0, v_1, \ldots, v_d$ be a sequence of vectors from $\mathbb{N}^d$. We say that ρ is a *base* if and only if $v_i(j) = 0$ for all $i, j \geq 1$ such that $i \neq j$. We use $first(\rho)$ to denote v_0. The ρ-*set*, denoted $\Theta(\rho)$, is defined by

$$\Theta(\rho) = \{v \in \mathbb{N}^d\colon v = v_0 + l_1 v_1 + \cdots + l_d v_d \text{ for some } l_1, \ldots, l_d \in \mathbb{N}\}.$$

Note that the ρ-set is a linear set (see, e.g., [24]). It is easy to see that each base is unique in the following sense.

Lemma 5.2. *If* ρ, ρ' *are bases such that* $\Theta(\rho) = \Theta(\rho')$, *then* $\rho = \rho'$.

Definition. Let $X \subseteq \Psi(\Sigma^*)$. We say that X is *periodic* if and only if there exists a base ρ such that $X = \Theta(\rho)$.

In view of Lemma 5.2, for each periodic $X \subseteq \Psi(\Sigma^*)$ there exists a unique base ρ such that $X = \Theta(\rho)$. In this case we say that ρ is the *base of* X and we write $\rho = base(X)$.

Definition. A language K is *periodic* if and only if K is commutative and $\Psi(K)$ is periodic. If K is periodic, then the base of $\Psi(K)$ is referred to as the *base of* K, denoted $base(K)$.

We have the following basic result for periodic languages.

Lemma 5.3. *Every periodic language is regular.*

Proof. Let K be a periodic language and let $base(K) = v_0, v_1, \ldots, v_d$. Clearly a word $w \in \Sigma^*$ is in K if and only if, for every $i \in \{1, \ldots, d\}$,

$$\#_{a_i}(w) \geq v_0(i) \quad \text{and} \quad \#_{a_i}(w) = v_0(i)(mod\, v_i(i)). \tag{2}$$

Here we follow the convention that $n(mod\ 0) = n$ for any $n \in \mathbb{N}$.

Consequently $K = K_1 \cap \cdots \cap K_d$ where $K_i = \{w \in \Sigma^*\colon (2) \text{ holds}\}$ for $1 \leq i \leq d$. It is easily seen that each K_i, $1 \leq i \leq d$, is regular and so K is regular. □

Two parameters of periodic languages, type and size, form a useful technical tool.

Definition. Let K be a periodic language where $base(K) = v_0, v_1, \ldots, v_d$.

(i) The *type of* K, denoted $type(K)$, is the pair of vectors (u_1, u_2) from $\mathbb{N}^d$ defined as follows:

$$u_1 = (v_0(1)(mod\, v_1(1)), \ldots, v_0(i)(mod\, v_i(i)), \ldots, v_0(d)(mod\, v_d(d)))$$

and

$$u_2 = (v_1(1), \ldots, v_i(i), \ldots, v_d(d)).$$

(ii) The *size of* K, denoted $size(K)$, is defined by

$$size(K) = \max_{1 \leq i \leq d} \{\max\{u_1(i), u_2(i)\}\} \quad \text{where } type(K) = (u_1, u_2).$$

Example. Let $\Sigma = \{a_1, a_2, a_3, a_4\}$ and let K be the periodic language such that

$$base(K) = (1, 6, 8, 0)\,(2, 0, 0, 0), (0, 3, 0, 0), (0, 0, 0, 0), (0, 0, 0, 7).$$

Then

$$type(K) = (u_1, u_2) \quad \text{where } u_1 = (1, 0, 8, 0) \text{ and } u_2 = (2, 3, 0, 7),$$

$$size(K) = \max\{2, 3, 8, 7\} = 8.$$

Lemma 5.4. *Let K_1, K_2 be periodic languages such that $type(K_1) = type(K_2)$. If $first(base(K_1)) \leq first(base(K_2))$, then $K_1 \supseteq K_2$.*

Proof. This is obvious. □

It turns out that well-quasi orders are naturally associated with certain families of periodic languages.

Definition. A family $\mathcal{F}$ of periodic languages is *bounded* if there exists a $q \in \mathbb{N}$ such that $size(K) \leq q$ for all $K \in \mathcal{F}$.

Theorem 5.5. *The containment relation* $\supseteq$ *is a well-quasi order on any bounded family of periodic languages.*

Proof. It is obvious that $\supseteq$ is a quasi order on any family of languages $\mathcal{F}$. Now assume that $\mathcal{F}$ is bounded family of periodic languages. Let $\{S_i\}$ be an infinite sequence of languages (possibly with repetitions) from $\mathcal{F}$. Since $\mathcal{F}$ is bounded, there exist only finitely many types for languages in $\mathcal{F}$ and thus we can find an infinite subsequence $\{T_i\}$ of $\{S_i\}$ such that $type(T_i) = type(T_j)$ for any $i, j \in \mathbb{N}$. Furthermore, within $\{T_i\}$ we can easily find an infinite subsequence $\{U_i\}$ such that $first(base(U_i)) \leq first(base(U_j))$ for any $i \leq j$. This is accomplished by first finding an infinite subsequence of $\{T_i\}$ in which the first components of the first vectors are increasing, and then choosing from this sequence one in which the second components are increasing and so on. By Lemma 5.4, for any $i \leq j$ we have $U_i \supseteq U_j$. Since $\{S_i\}$ was chosen as an arbitrary sequence, $\supseteq$ is a well-quasi order on $\mathcal{F}$ (see definition (iii) of a well-quasi order in Section 3). □

The above theorem immediately yields a type of 'compactness' result for sets covered by bounded families of periodic languages.

Corollary 5.6. *Let $\mathcal{F}$ be any bounded family of periodic languages. Then there exists a finite $\mathcal{S} \subseteq \mathcal{F}$ such that $\bigcup_{K \in \mathcal{F}} K = \bigcup_{K \in \mathcal{S}} K$.*

Proof. This follows directly from Theorem 5.5 (see definition (iv) of well-quasi order in Section 3). □

Since any finite set of periodic languages is a bounded family, the following result generalizes our previous regularity result for periodic languages (Lemma 5.3).

Corollary 5.7. *For any bounded family of periodic languages $\mathcal{F}$, $\bigcup \mathcal{F}$ is regular.*

Proof. This follows directly from Corollary 5.6 and Lemma 5.3 (because the union of any finite set of regular languages is regular). □

6. Commutative linear languages

In this section we prove that each commutative linear language is regular. This is accomplished by providing a representation of commutative linear languages in terms of periodic languages.

Lemma 6.1. *For any commutative linear language K, there exists a $q \in \mathbb{N}^+$ such that for every $w \in K$ there exists a periodic language $L_w \subseteq K$ where $w \in L_w$ and size$(L_w) \leq q$.*

Proof. Let K be a commutative linear language and let $G = (\Omega, \Sigma, P, S)$ be a linear grammar generating K where Ω is its total alphabet, Σ its terminal alphabet, P its set of productions and S its axiom. Clearly we can assume that each production of G is in one of the following three forms:

$A \to Ba$, $A \to aB$ and $A \to a$ where A, B are nonterminals ($A, B \in \Omega - \Sigma$) and a is a terminal ($a \in \Sigma$).

Let $m = \#\Omega$. We define the sequence $\{q_i\}_{i \geq 1}$ of positive integers as follows: $q_1 = m + 1$ and $q_{i+1} = (q_1 + \cdots + q_i + 1)(m + 1)$ for $i \geq 1$.

Then we set $q = 2q_m$.

Let $w \in K$. Let $\rho = v_0, v_1, \ldots, v_d$ be the base defined as follows:

$$v_0 = \Psi(w), \qquad \text{if } 1 \leq i \leq d \text{ is such that } v_0(i) \leq q, \text{ then } v_i(i) = 0.$$

If, for every $i \in \{1, \ldots, d\}$, $v_0(i) \leq q$, then all components of ρ are defined and we are done. Otherwise we proceed as follows.

Let $\{b_1, \ldots, b_s\}$ be all the letters from *alph*(w) such that $\#_{b_j}(w) > q$ for $1 \leq j \leq s$. Now let $w' = b_1^{q_1} \cdots b_s^{q_s} u b_s^{q_s} \cdots b_1^{q_1}$ where u is a fixed word such that $b_1^{q_1} \cdots b_s^{q_s} u b_s^{q_s} \cdots b_1^{q_1} \in com(w)$. Since $q = 2q_m$, w' is well defined. For $1 \leq i \leq s$ we refer to the leftmost occurrence of $b_i^{q_i}$ in w' as the *left i-block* and the rightmost occurrence as the *right i-block*; the left i-block together with the right i-block form the *i-block* of w'.

Consider a derivation tree D of w' in G; the path of D originating in its root and ending on a leaf of D such that the direct ancestor of the last node (the leaf) has one descendant only is called the *spine* of D and denoted τ. A sequence of consecutive nodes of τ is called a *segment* (of τ). The label of a node e of τ is denoted by $l(e)$. If $\mu = e_1 \cdots e_k e_{k+1}$ is a segment of τ such that $k \geq 1$, $e_1, \ldots, e_{k+1}$ are nodes of τ, $l(e_1) = l(e_{k+1})$ and $l(e_j) \neq l(e_1)$ for $2 \leq j \leq k$, then μ is called a ***repeat*** (of τ); $e_1 \cdots e_k$ is the *front* of μ (denoted *front*(μ)). The ***contribution*** of a segment μ of τ are the occurrences in w' which are 'derived' from nodes of μ (in other words those occurrences in w' which have ancestors among the nodes of μ).

The following technical result is very crucial to our proof of Lemma 6.1.

Claim 6.2. *For every $1 \leq i \leq s$ there exists a repeat μ on τ such that the contribution of front(μ) is contained in the i-block of w'.*

Proof of Claim 6.2. The proof goes by induction on i, $1 \leq i \leq s$. Let $i = 1$.

Consider the segment of τ consisting of its first $(m + 1)$ nodes. Since $q_1 = m + 1$, it is clear that this segment contributes only to the first block of w'. On the other hand, the length of this segment is $(m+1)$ and so it must contain a repeat. Hence the claim holds for $i=1$.

Assume that the claim holds up to the $(i-1)$-block of w' where $2 \leqslant i \leqslant s$. We will demonstrate now that it holds for the i-block of w'.

Let U be the rightmost occurrence of b_{i-1} in the left $(i-1)$-block in w' and let T be the leftmost occurrence of b_{i-1} in the right $(i-1)$-block of w'. Let O_U be the ancestor of U on τ and let O_T be the ancestor of T on τ.

Thus we have the situation diagramed in Fig. 6.1. (We have assumed that O_U is closer to the root than O_T; clearly we can assume this without loss of generality.)

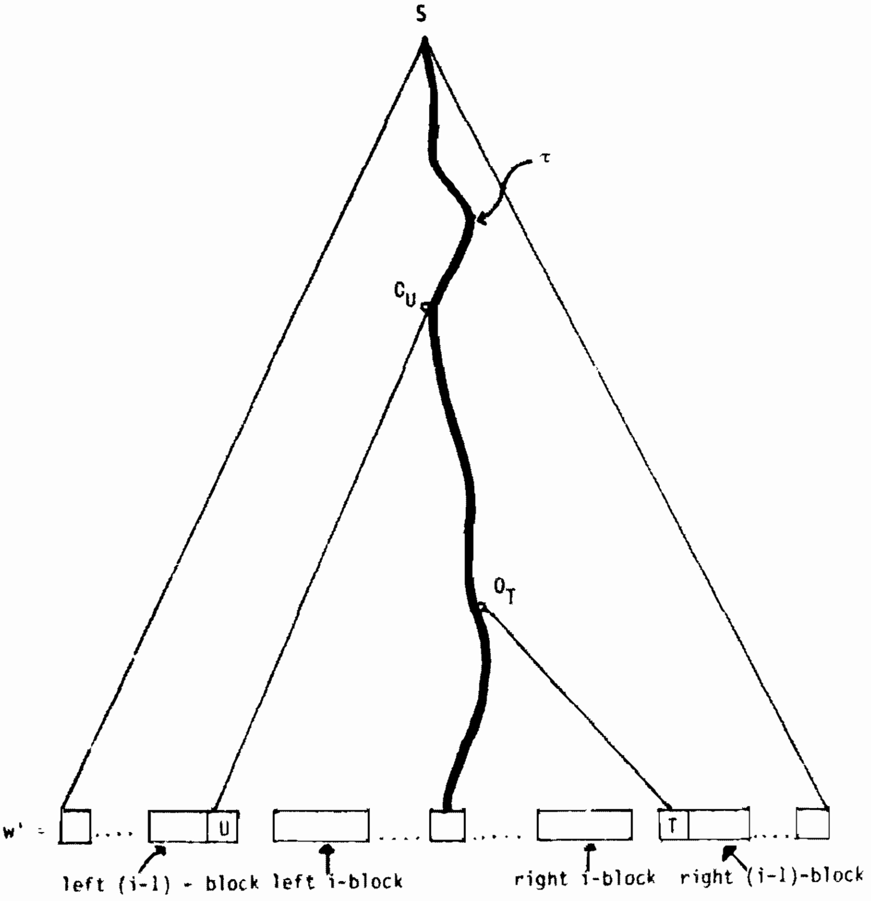

Fig. 6.1.

Clearly all nodes above O_U contribute either to the left of U or to the right of T. Now let $Q_1, \ldots, Q_l$ be *all* the nodes strictly between O_U and O_T such that they contribute to the right of T.

Since $|b_1^{q_1}b_2^{q_2}\cdots b_{i-2}^{q_{i-2}}b_{i-1}^{q_{i-1}}| = q_1 + \cdots + q_{i-1}$, we clearly have

$$l+1 \leqslant q_1 + \cdots + q_{i-1}. \tag{3}$$

Now let $z_1, \ldots, z_l, z_{l+1}$ be segments of τ defined as follows:

z_1 consists of all the nodes strictly between O_U and Q_1,

z_2 consists of all the nodes strictly between Q_1 and Q_2,

.
.
.

z_l consists of all the nodes strictly between Q_{l-1} and Q_l,

z_{l+1} consists of all the nodes strictly between Q_l andO_T.

We now consider two cases.

Case 1. At least one of the segments $z_1, \ldots, z_l$ consists of more than m nodes.

Let i_0 be the smallest index j such that z_j consists of more than m nodes. In z_{i_0} we consider the segment γ consisting of the first $m+1$ nodes. Clearly, this segment contains a repeat, say μ. Note that all the nodes from $z_1, z_2, \ldots, z_{i_0-1}, \gamma$ contribute to the right of U (but to the left of T). The number of occurrences contributed to w' by all the nodes from $z_1, \ldots, z_{i_0-1}, \gamma$ is not greater than $(l+1)(m+1)$ and so by (3) it is not greater than $(q_1+\cdots+q_{l-1}+1)(m+1)$. Since the length of the left and the right i-block equals q_i, this means that all occurrences contributed by nodes from $z_1, \ldots, z_{i_0-1}, \gamma$ are within the i-block. Thus in this case the claim holds for the ith block.

Case 2. Each of the segments $z_1, \ldots, z_{l+1}$ consists of no more than m nodes.

Clearly in this case the number of occurrences contributed to w' by all the nodes from $z_1, \ldots, z_{l+1}$ does not exceed $(l+1)m$ and (because the length of the left and right i-block is q_i) all of these occurrences are within the i-block. Moreover, from (3) and from the definition of q_i it follows that if we consider the segment ρ of τ consisting of $(m+1)$ nodes immediately following O_T, then all the nodes from ρ will contribute to the i-block of w'. But ρ must contain a repeat and so also in this case the claim holds for the ith block.

Hence we have completed the induction and the claim holds. □

Proof of Lemma 6.1 (*continued*). Now that the claim is proved we complete the definition of ρ as follows.

For each $i \in \{1, \ldots, s\}$ let $k(b_i)$ be the length of the front of the repeat μ on τ which satisfies the statement of Claim 6.2 and has the shortest length. If $b_i = a_j$ for $1 \leq j \leq d$, we set $v_i(j) = k(b_i)$. Thus ρ is now completely defined: $\rho = v_0, v_1, \ldots, v_d$.

We set $L_{w'} = L(\Theta(\rho))$. In order to show that $L_{w'} \subseteq K$, it suffices to show that $\Theta(\rho) \subseteq \Psi(K)$ (see Lemma 5.1). Let $v \in \Theta(\rho)$, hence $v = v_0 + l_1 v_1 + \cdots + l_d v_d$ where $l_1, \ldots, l_d \in \mathbb{N}$. If $v_i(i) \neq 0$ for $1 \leq i \leq d$, then in the derivation tree D of w' (from the proof of the above claim) we will 'iterate' l_i times a repeat of the length $k(a_i)$ contributing to the i-block (and we do it for each i satisfying $v_i(i) \neq 0$). In this way we get the word $w'(l_1, \ldots, l_d)$ such that $\Psi(w'(l_1, \ldots, l_d)) = v$. Thus $v \in \Psi(K)$.

Consequently $\Theta(p) \subseteq \Psi(K)$ and so $L_{w'} \subseteq K$. Clearly *size* $(L_{w'}) \leq q$. Finally we notice that $w \in L_{w'}$ (because $w' \in com(w)$) and so if we set $L_w = L_{w'}$, the lemma holds. □

Theorem 6.3. *A language is a commutative linear language if and only if it is a finite union of periodic languages.*

Proof. Assume that K is a finite union of periodic languages. Then, by Lemma 5.3, K is a commutative regular language and so a commutative linear language. On the other hand, if K is a commutative linear language, then using Lemma 6.1 we can find a bounded family $\mathcal{F}$ of periodic languages such that $K = \bigcup \mathcal{F}$. Thus, by Corollary 5.6, K is a finite union of periodic languages. □

The following corollary of Theorem 6.3 gives a partial answer to a conjecture from [18].

Corollary 6.4. *If K is a commutative linear language, then K is regular.*

Proof. Follows from Theorem 6.3 and Lemma 5.3. □

In [18] Latteux conjectures that the above result holds for commutative quasi-rational languages (the family of quasi-rational languages is the substitution closure of the family of linear languages).

Furthermore, from our results also the following theorems follows. (One should note here that this theorem follows also from well-known results concerning bounded languages.)

Theorem 6.5. *For any commutative language K, the following are equivalent:*

(i) *K is regular.*

(ii) *K is the union of a finite set of periodic languages.*

(iii) $K = \cup \mathcal{F}$ *for some bounded family of periodic languages* $\mathcal{F}$.

Proof. From Corollary 5.6 we have that (iii) implies (ii) and from Lemma 5.3 we know that (ii) implies (i). Finally, the fact that (i) implies (iii) follows from Theorem 6.3, since any regular language is linear. □

7. Discussion

In our paper we have presented some conditions enforcing regularity of context-free languages. Additional results in this direction are given in [10]. While the primary fruits of the present investigation lie in our two main results (Theorems 4.12 and 6.3) we hope that the characterizations of regularity for commutative languages (Theorem 6.5) and for languages in general (Theorem 3.3) will also prove useful.

It should be noted that with each of the latter theorems we have two equivalent characterizations of regularity. The first is a simpler, 'finitistic' characterization (i.e., finite unions of periodic languages in Theorem 6.5 and unions of equivalence classes from congruences of finite index in the Myhill–Nerode result). The second is a generalized form of the first, achieved through the application of well-quasi order theory. (Thus we have the unions of bounded families of periodic languages in Theorem 6.5 and the closed sets induced by monotone well-quasi orders in the generalized Myhill–Nerode result.) While characterizations of the first kind are useful as 'normal form' results, the characterizations of the second kind, because of their greater generality, are more useful in proving specific languages to be regular. This is especially the case when those 'finitistic' aspects of the language which make it regular are only found deeply hidden in its structure.

The use of the theory of well-quasi orders in this manner brings up the question of effectiveness. Suppose we prove a language to be regular using a characterization of the second kind, i.e., using well-quasi order theory. How large is the smallest 'normal form' representation of this language? This is a particularly difficult problem for the generalized Myhill–Nerode theorem. Let the *syntactic complexity* of a regular set be the smallest index of any congruence which represents it as a union of equivalence classes. Are there any parameters that can be given for a monotone well-quasi order which determine the syntactic complexity of its closed sets? In particular, what is the syntactic complexity of the pure unitary language $cl_{\leq_I}(\lambda)$ when I is subword unavoidable, with subword avoidance bound k_0? At present we have no answers to these questions.

It is also natural to ask whether these results can be used to explore the issue of regularity in classes other than those we have investigated. In particular, for what other classes of languages does requiring commutativity imply regularity? Can we use Theorem 6.5 here? Work on this question is in progress and we hope to report on it soon. On the other hand, the derivation relation is a monotone quasi order in many of the types of grammatically defined language classes that have been studied in the area of formal language theory. Included in this category are the scattered context languages [12] and the languages generated by various forms of matrix grammars (see, e.g., [24]), as well as the various classes of languages generated by grammars based on normal semi-Thue derivations. The significance of the generalized Myhill–Nerode theorem with respect to the study of regularity in these language classes has yet to be explored.

Furthermore, in applying this theorem, in each case we would like to know under what conditions the given derivation relation is a well-quasi order. Here we would like to have a set of general results concerning monotone well-quasi orders on freely generated monoids. Our Lemmas 4.4 and 4.5 are only very rudimentary results in this direction. As a specific example, let us consider the derivation relation implicit in the context-free languages. This relation is represented by a semi-Thue system T where $\langle u, v\rangle \in T$ implies that $|u| = 1$. When does a system of this type generate a well-quasi order? Suppose we also allow $|u| = 0$, i.e., $u = \lambda$. Such a system would be called a *monadic* semi-Thue system (see [4]). Can we generalize the Higman theorem further to give a characterization monadic semi-Thue systems which generate well-quasi orders? Or perhaps even arbitrary semi-Thue systems?

Finally, one can also consider some of the standard language theoretic questions for the class of unitary languages. One such problem, that of the emptiness of intersection, is solved for this class in [15]. There it is shown that it is undecidable whether or not two pure unitary languages have a non-λ intersection (i.e., an intersection other than $\{\lambda\}$). This is shown even in the special case in which $S(I)$ is Church–Rosser for each insertion set (see, e.g., [6]). However, another basic question remains open. Is it decidable whether or not two unitary languages are equivalent? We might note here that using the techniques from [9], this can be construed as a subproblem of the equivalence problem for D0S languages.

Acknowledgment

The authors are indebted to H.C.M. Kleijn and R. Verraedt for comments concerning the first draft of this paper.

References

[1] S. Adian. Defining relations and algorithmic problems for groups and semigroups, *Proc. Steklov Institute Math.* **85** (1966) (English version published by American Math. Soc., 1967).

[2] J.M. Autebert, J. Beauquier, L. Boasson and M. Latteux, Very small families of algebraic nonrational languages, in: R. Book, ed., *Formal Language Theory* (Academic Press, London–New York, 1981).

[3] J.M. Autebert, J. Beauquier, L. Boasson and M. Nivat, Quelques problèmes ouverts en théorie des languages algébraiques, *RAIRO Informatique Theorique* **13** (1979) 363–379.

[4] R. Book, M. Jantzen and C. Wrathall, Monadic Thue systems, *Theoret. Comput. Sci.* **19** (1982) 231–251.

[5] N. Chomsky and M. Schutzenberger, The algebraic theory of context-free languages, in: Braffort and Hirshberg, eds., *Computer Programming and Formal Systems* (North-Holland, Amsterdam, 1963) pp. 118–161.

[6] Y. Cochet, Church-Rosser congruences on free semigroups, *Colloq. Math. Soc. Janos Bolyai: Algebraic theory of semigroups* **20** (1976) 51–60.

[7] Y. Cochet and M. Nivat, Une généralisation des ensembles de Dyck, *Israel J. Math.* **9** (1971) 389–395.

[8] J.H. Conway, *Regular Algebra and Finite Machines* (Chapman & Hall, London, 1971) pp. 63–64.

[9] A. Ehrenfeucht and G. Rozenberg, On basic properties of D0S systems, *Inform. and Control* **47** (1980) 137–153.

[10] A. Ehrenfeucht, G. Rozenberg and D. Haussler, Conditions enforcing regularity of context-free languages, *Lecture Notes in Computer Science* (Springer, Berlin, 1982).

[11] P. Erdös and R. Rado, Sets having the divisor property, solution to problem 4358, *Amer. Math. Monthly* **59** (1952) 255–257.

[12] S. Greibach and J. Hopcroft, Scattered context grammars, *J. Comput. Systems Sci.* **3** (3) (1969) 233–247.

[13] L.H. Hains, On free monoids partially ordered by embedding, *J. Combin. Theory* **6** (1969) 94–98.

[14] M.A. Harrison, *Introduction to Formal Language Theory* (Addison-Wesley, Reading, MA, 1978).

[15] D. Haussler, Insertion languages, *Inform. Sci.*, to appear.

[16] G. Higman, Ordering by divisibility in abstract algebras, *Proc. London Math. Soc.* **3** (2) (1952) 326–336.

[17] J.B. Kruskal, The theory of well-quasi ordering: A frequently discovered concept, *J. Combin. Theory* (*Ser. A*) **13** (1972) 297–305.

[18] M. Latteux, Cônes rationnelles commutatifs, *J. Comput. and Systems Sci.* **18** (1979) 307–333.

[19] R. Laver, Well-quasi orderings of sets of finite sequences, *Math. Proc. Cambridge Philos. Soc.* **79** (1976) 1–10.

[20] J. Myhill, Finite automata and the representation of events, Wright Air Development Command Tech. Rept. 57-624 (1957)112–137.

[21] C. St. J.A. Nash-Williams, On well-quasi-ordering finite trees, *Proc. Cambridge Phil. Soc.* **59** (1963) 833–835.

[22] A. Nerode, Linear automaton transformations, *Proc. Amer. Math. Soc.* **9** (1958) 541–544.

[23] E.L. Post, Recursive unsolvability of a problem of Thue, *J. Symb. Logic* **12** (1947) 1–11.

[24] A. Salomaa, *Formal Languages* (Academic Press, New York, 1973).

[25] A. Thue, Probleme über veranderungen von zeichnenreihen nach gegebenen regeln, *Skr. Vidensk. Selsk.* **1** (10) (1914).