Available online at www.sciencedirect.com
SCIENCE DIRECT®

Theoretical Computer Science 327 (2004) 255–268

Theoretical Computer Science

www.elsevier.com/locate/tcs

Well quasi-orders and context-free grammars[☆,☆☆]

Flavio D'Alessandro[a], Stefano Varricchio[b,*]

[a] *Dipartimento di Matematica, Università di Roma "La Sapienza" Piazzale Aldo Moro 2, 00185 Roma, Italy*
[b] *Dipartimento di Matematica, Università di Roma "Tor Vergata", Viale della Ricerca Scientifica 1, 00133 Roma, Italy*

Received 3 October 2003; accepted 31 March 2004

Abstract

Let G be a context-free grammar and let L be the language of all the words derived from any variable of G. We prove the following generalization of Higman's theorem: any division order on L is a well quasi-order on L. We also give applications of this result to some quasi-orders associated with unitary grammars.

Keywords: Well quasi-orders; Context-free grammars; Unavoidable sets of words

1. Introduction

A *quasi-order* on a set S is called a *well quasi-order* (*wqo*) if every non-empty subset X of S has at least one minimal element in X but no more than a finite number of (non-equivalent) minimal elements.

Well quasi-orders have been widely investigated in the past. In [9] Higman gives a very general theorem on division orders in abstract algebras that in the case of semigroups becomes: *Let S be a semigroup quasi-ordered by a division order $\leqslant$. If there exists a generating set of S well quasi-ordered by $\leqslant$, then S will also be so.* From this one derives that the *subsequence ordering* in free monoids is a wqo.

☆ This work was partially supported by MIUR project "Linguaggi formali e automi: teoria e applicazioni".
☆☆ A short version of this paper appeared in the proceedings of the conference DLT03.

* Corresponding author.
E-mail addresses: dalessan@mat.uniroma1.it (F. D'Alessandro), varricch@mat.uniroma2.it (S. Varricchio)
URLs: http://www.mat.uniroma1.it/people/dalessandro/, http://axp.mat.uniroma2.it/~varricch/

doi:10.1016/j.tcs.2004.03.069

In [12] Kruskal extends Higman's result, proving that certain embeddings on finite trees are well quasi-orders. In the last years many papers have been devoted to the applications of wqo's to formal language theory. The most important result is a generalization of the famous Myhill–Nerode theorem on regular languages. In [6] Ehrenfeucht et al. proved that a language is regular if and only if it is upward-closed with respect to a monotone well quasi-order. From this result many regularity conditions have been derived (see for instance [2–5]).

In [6] unavoidable sets of words are characterized in terms of the wqo property of a suitable unitary grammar: a set I is unavoidable if and only if the derivation relation $\Rightarrow_I^*$ of the unitary semi-Thue system associated with the finite set $I \subseteq A^+$ is a wqo. An extension of the previous result has been given by Haussler in [8], considering set of words which are *subsequence unavoidable*.

In [11] some extensions of Higman and Kruskal's theorem to regular languages and rational trees have been given. Further applications of the wqo theory to formal languages are given in [7,10].

In this paper we give a new generalization of Higman's theorem. First of all we define the notion of *division order* on a language L: a quasi order $\leqslant$ on A^* is called a *division order* on L if it is monotone and for any $u, v \in L$ if u is factor of v then $u \leqslant v$. When L is the whole free monoid A^* this notion is equivalent to the classical one, but, in general, a quasi-order on A^* could be a division order on a set L and not on A^*. Then, given a context-free grammar G with set of variables $V = \{X_1, X_2, \ldots, X_n\}$, let L_i be the language of the words generated setting X_i as start symbol and let $L = \bigcup_{i=1}^n L_i$. Our main theorem states that any division order on L is a well quasi-order on L. In particular, if L is a context-free language generated by a grammar with only one variable, then any division order on L is a wqo on L. This generalizes Higman's theorem on finitely generated free monoids, since for any finite alphabet A, the set A^* can be generated by a context-free grammar having only one variable. We also introduce the notion of *weak division order* on a language and we extend the previous result, under the additional hypothesis that $\varepsilon \in L_i$ for any i.

In the second part of the paper we study the wqo property in relation to some quasi-orders associated with unitary grammars. Let I be a finite set of words and let $\Rightarrow_I^*$ be the derivation relation associated with the semi-Thue system

$$\{\varepsilon \to u,\ u \in I\}.$$

One can also consider the relation $\vdash_I^*$ as the transitive and reflexive closure of $\vdash_I$ where $v \vdash_I w$ if

$$v = v_1 v_2 \cdots v_{n+1},$$

$$w = v_1 a_1 v_2 a_2 \cdots v_n a_n v_{n+1},$$

where the a_i's are letters, and $a_1 a_2 \cdots a_n \in I$.

We set $L_I^\varepsilon = \{w \in A^* \mid \varepsilon \Rightarrow_I^* w\}$, $L_{\vdash_I}^\varepsilon = \{w \in A^* \mid \varepsilon \vdash_I^* w\}$ and prove that

- There exists a finite set I such that $\Rightarrow_I^*$ is not a wqo on L_I^ε;
- There exists a finite set I such that $\vdash_I^*$ is not a wqo on $L_{\vdash_I}^\varepsilon$;
- For any finite set I the relation $\vdash_I^*$ is a wqo on L_I^ε.

2. Preliminaries

The main notions and results concerning quasi-orders and languages are shortly recalled in this section. Let A be a finite *alphabet* and A^* the free monoid generated by A. The elements of A are usually called *letters* and those of A^* *words*. The identity of A^* is denoted ε and called the *empty word*.

A nonempty word $w \in A^*$ can be written uniquely as a sequence of letters as $w = a_1a_2 \cdots a_n$, with $a_i \in A$, $1 \leqslant i \leqslant n$, $n > 0$. The integer n is called the *length* of w and denoted $|w|$. For all $a \in A$, $|w|_a$ denotes the number of occurrences of the letter a in w. Let $w \in A^*$. The word $u \in A^*$ is a *factor* of w if there exist $p, q \in A^*$ such that $w = puq$. If $w = uq$, for some $q \in A^*$ (resp. $w = pu$, for some $p \in A^*$), then u is called a *prefix* (resp. a *suffix*) of w. The set of all prefixes (resp. suffixes, factors) of w is denoted $Pref(w)$ (resp. $Suf(w)$, $Fact(w)$). A word u is a *subsequence* of a word v if $u = a_1a_2 \cdots a_n$, $v = v_1a_1v_2a_2 \cdots v_na_nv_{n+1}$ with $a_i \in A$, $v_i \in A^*$.

A subset L of A^* is called a *language*. If L is a language of A^*, then $\mathrm{alph}(L)$ is the smallest subset B of A such that $L \subseteq B^*$. A binary relation $\leqslant$ on a set S is a *quasi-order* (qo) if $\leqslant$ is reflexive and transitive. Moreover, if $\leqslant$ is symmetric, then $\leqslant$ is an equivalence relation. The meet $\leqslant \cap \leqslant^{-1}$ is an equivalence relation $\sim$ and the quotient of S by $\sim$ is a *poset* (partially ordered set).

An element $s \in X \subseteq S$ is *minimal* in X with respect to $\leqslant$ if, for every $x \in X$, $x \leqslant s$ implies $x \sim s$. For $s, t \in S$ if $s \leqslant t$ and s is not equivalent to t mod $\sim$, then we set $s < t$. A part X of S is *upper-closed*, or simply *closed*, with respect to $\leqslant$ if the following condition is satisfied:

$$\text{if } \ x \in X \ \text{ and } \ x \leqslant y \ \text{ then } \ y \in X.$$

We shall denote by $\mathrm{Cl}(X)$ the *closure* of X,

$$\mathrm{Cl}(X) = \{s \in S \mid \exists\, x \in X \ \text{ such that } \ x \leqslant s\},$$

so that X is closed if and only if $X = \mathrm{Cl}(X)$. For any $X \subseteq S$ one has $X \subseteq \mathrm{Cl}(X)$. Moreover, if $Y \subseteq X$, then $\mathrm{Cl}(Y) \subseteq \mathrm{Cl}(X)$. A closed set X is called *finitely generated* if there exists a finite subset F of X such that $\mathrm{Cl}(F) = X$.

A quasi-order in S is called a *well quasi-order* (wqo) if every non-empty subset X of S has at least one minimal element but no more than a finite number of (non-equivalent) minimal elements. We say that a set S is *well quasi-ordered* (wqo) by $\leqslant$, if $\leqslant$ is a well quasi-order on S.

There exists several conditions which characterize the concept of well quasi-order and that can be assumed as equivalent definitions (cf. [5]).

Theorem 1. *Let S be a set quasi-ordered by $\leqslant$. The following conditions are equivalent*:
- (i) *$\leqslant$ is a well quasi-order*;
- (ii) *the ascending chain condition holds for the closed subsets of S*;
- (iii) *every infinite sequence of elements of S has an infinite ascending subsequence*;
- (iv) *if $s_1, s_2, \ldots, s_n, \ldots$ is an infinite sequence of elements of S, then there exist integers i, j such that $i < j$ and $s_i \leqslant s_j$*;

(v) *there exists neither an infinite strictly descending sequence in S* (*i.e.* $\leqslant$ *is well founded*), *nor an infinity of mutually incomparable elements of S*;
(vi) *S has the finite basis property*, *i.e. every closed subset of S is finitely generated.*

Let $\sigma = \{s_i\}_{i \geqslant 1}$ be an infinite sequence of elements of *S*. Then σ is called *good* if it satisfies condition iv of Theorem 1 and it is called *bad* otherwise, that is, for all integers i, j such that $i < j$, $s_i \nleqslant s_j$. It is worth noting that, by condition iv above, a useful technique to prove that $\leqslant$ is a wqo on *S* is to prove that no bad sequence exists in *S*.

If ρ and σ are two relations on sets *S* and *T*, respectively, then the direct product $\rho \otimes \sigma$ is the relation on $S \times T$ defined as

$$(a, b)\ \rho \otimes \sigma\ (c, d) \iff a\ \rho\ c \ \text{ and }\ b\ \sigma\ d.$$

The following lemma is well known (*see* [5, Chap. 6]).

Lemma 1. *The following conditions hold*:
(i) *Every subset of a wqo set is wqo*;
(ii) *If S and T are wqo by* $\leqslant_S$ *and* $\leqslant_T$, *respectively, then* $S \times T$ *is wqo by* $\leqslant_S \otimes \leqslant_T$.

Let us now suppose that the set *S* is a semigroup. Let $S^1 = S$ if *S* is a monoid, otherwise S^1 is the monoid obtained by adding the identity to *S*.

Definition 1. A quasi-order $\leqslant$ in a semigroup *S* is *monotone on the right* (*resp. on the left*) if for all $x_1, x_2, y \in S$

$$x_1 \leqslant x_2 \ \text{ implies }\ x_1 y \leqslant x_2 y \ \ (\text{resp. } yx_1 \leqslant yx_2).$$

A quasi-order is *monotone* if it is monotone on the right and on the left.

Definition 2. A quasi-order $\leqslant$ in a semigroup *S* is a *division order* if it is monotone and, for all $s \in S$ and $x, y \in S^1$,

$$s \leqslant xsy.$$

The ordering by division in abstract algebras was studied by Higman [9] who proved a general theorem that in the case of semigroups becomes:

Theorem 2. *Let S be a semigroup quasi-ordered by a division order* $\leqslant$. *If there exists a generating set of S well quasi-ordered by* $\leqslant$ *then so will be S.*

If *n* is a positive integer, then the set of all positive integers less or equal than *n* is denoted $[n]$. If *f* is a map then $\mathrm{Im}(f)$ denotes the set of the images of *f*.

3. Main result

In this section we prove our main result. For this purpose, it is useful to give some preliminary definitions and results. We assume the reader to be familiar with the basic theory of context–free languages. It is useful to recall few elements of the vocabulary (cf. [1]).

A *context-free grammar* is a triplet $G = (V, A, P)$ where V and A are finite sets of *variables* and *terminals*, respectively. P is the set of *productions*: each element of P is of the form $X \to u$ with $X \in V$ and $u \in \{V \cup A\}^*$.

The relation $\Rightarrow_G$, simply denoted by $\Rightarrow$, is the binary relation on the set $\{V \cup A\}^*$ defined as: $w_1 \Rightarrow w_2$ if and only if $w_1 = w'Xw''$, $w_2 = w'uw''$ where $X \to u$ is a production of G and $w', w'' \in \{V \cup A\}^*$. The relation $\Rightarrow^*$ is the reflexive and transitive closure of $\Rightarrow$. Let $V = \{X_1, X_2, \ldots, X_n\}$. For every $i = 1, \ldots, n$, the language generated by X_i is $L(X_i) = \{u \in A^* \mid X_i \Rightarrow^* u\}$. We shall adopt the convention to denote $L(X_i)$ by L_i whenever no ambiguity or confusion arises.

Definition 3. Let $\leqslant$ be a quasi-order on A^*. Then $\leqslant$ is said to be *compatible* with G if the following condition holds: for every production of G of the kind $X_i \longrightarrow u_1Y_1u_2Y_2 \cdots u_mY_mu_{m+1}$, where $u_k \in A^*$, for $k = 1, \ldots, m+1$, and $Y_k \in V$, $k = 1, \ldots, m$, one has

$$x_k \leqslant u_1x_1u_2x_2 \cdots u_mx_mu_{m+1},$$

for any choice of $x_i \in L(Y_i)$, for $i = 1, \ldots, m$ and for any $k \in \{1, \ldots, m\}$.

The following result holds.

Proposition 1. *If $\leqslant$ is a monotone quasi-order compatible with G, then $\leqslant$ is a wqo on $L = \bigcup_{i=1}^n L_i$.*

Proof. In this proof, for the sake of simplicity, we assume that the grammar G contains neither unitary productions nor ε-productions. The proof is by contradiction. Hence there exists a bad sequence in L. Following an idea of Nash–Williams (see [12]), we construct a bad sequence $\gamma = \{v_i\}_{i \geqslant 1}$ in L, which is "minimal" in the sense we shall explain later.

Select $v_1 \in L$ such that v_1 is the first term of a bad sequence in L and its length $|v_1|$ is as small as possible.

Suppose, by induction, that we have constructed the elements $v_1, \ldots, v_{n-1}$ of γ such that there is a bad sequence of L whose first $n-1$ elements are $v_1, \ldots, v_{n-1}$. Then select a word $v_n \in L$ such that $v_1, \ldots, v_{n-1}, v_n$ (in that order) are the first n elements of a bad sequence in L and $|v_n|$ is as small as possible. This construction yields a bad sequence $\gamma = \{v_i\}_{i \geqslant 1}$ in L. This sequence is minimal in the following sense: let $\alpha = \{z_i\}_{i \geqslant 1}$ be a bad sequence of L and let k be a positive integer such that, for $i = 1, \ldots, k$, $z_i = v_i$, then $|v_{k+1}| \leqslant |z_{k+1}|$.

Since the set of productions P is finite, we may consider a subsequence $\sigma = \{v_{i_\ell}\}_{i_\ell \geqslant 1}$ of the sequence above, which satisfies the following property:

$$\forall\, \ell \geqslant 1, \quad X_k \Rightarrow\ p \Rightarrow^* v_{i_\ell}, \tag{1}$$

where $X_k \to p$ is a production and $p = u_1Y_1u_2Y_2 \cdots u_mY_mu_{m+1}$. By the sake of simplicity, let us rename the terms of σ as: for every $\ell \geqslant 1$, $w_\ell = v_{i_\ell}$. Hence, by Eq. (1), for every

$\ell \geqslant 1$, one has

$$w_\ell = u_1 x_1^\ell u_2 x_2^\ell \cdots u_m x_m^\ell u_{m+1},$$
$$\text{with } x_1^\ell \in L(Y_1),\ x_2^\ell \in L(Y_2),\ \ldots,\ x_m^\ell \in L(Y_m).$$

For every $j = 1, \ldots, m$, set $F_j = \{x_j^i\}_{i \geqslant 1}$. The following claim is crucial.

Claim. *For every $j = 1, \ldots, m$, F_j is well quasi-ordered by $\leqslant$.*

Proof of the Claim. By contradiction, let j be a positive integer with $1 \leqslant j \leqslant m$ such that F_j is not well quasi-ordered by $\leqslant$. Let $\tau = \{y_i\}_{i \geqslant 1}$ be a bad sequence in F_j.

We first observe that, for all $i \geqslant 1$, there exists a positive integer $g(i)$ such that $y_i = x_j^{g(i)}$. Without loss of generality we may assume that for every $i \geqslant 1$, $g(i) \geqslant g(1)$. Indeed, if the above condition is not satisfied one can consider a subsequence of τ satisfying this property.

Consider now the sequence

$$v_1,\ v_2,\ \ldots,\ v_{i_{g(1)}-1},\quad y_1,\ y_2,\ \ldots,\ y_i\ \ldots .$$

By construction, every term of the above sequence belongs to L. Moreover one easily proves the latter sequence is bad. Since γ and $\{y_i\}_{i \geqslant 1}$ are bad sequences in L, this amounts to show that for h, k, $1 \leqslant h \leqslant i_{g(1)} - 1$, $k \geqslant 1$, one has $v_h \not\leqslant y_k$. Indeed, suppose $v_h \leqslant y_k$. Since $y_k = x_j^{g(k)}$, then $v_h \leqslant x_j^{g(k)}$. Since for every $\ell = 1, \ldots, m$, $x_\ell^{g(k)} \in L(Y_\ell)$, the fact that $\leqslant$ is compatible with G entails

$$x_j^{g(k)} \leqslant u_1 x_1^{g(k)} u_2 \cdots u_m x_m^{g(k)} u_{m+1} = w_{g(k)} = v_{i_{g(k)}}.$$

Hence $v_h \leqslant v_{i_{g(k)}}$. Since $g(1) \leqslant g(k)$, one has $h < i_{g(1)} \leqslant i_{g(k)}$ and this contradicts that γ is bad. Hence $v_h \not\leqslant y_k$.

Now we observe that y_1 is a proper factor of $w_{g(1)} = v_{i_{g(1)}}$, since the grammar contains neither unitary productions nor ε-productions. Thus $|y_1| < |v_{i_{g(1)}}|$ and this contradicts that γ is minimal. Hence, no bad sequence in F_j exists and so F_j is well quasi-ordered by $\leqslant$. $\diamond$

Let $\mathcal{F} = F_1 \times F_2 \times \cdots \times F_m$. By condition (ii) of Lemma 1 and the claim above, one has that the set $\mathcal{F}$ is well quasi-ordered by the canonical extension of $\leqslant$ on $\mathcal{F}$. Consider now the sequence of $\mathcal{F}$ defined as

$$\{\ (x_1^i,\ x_2^i,\ x_3^i,\ \ldots,\ x_m^i)\}_{i \geqslant 1}.$$

Since $\mathcal{F}$ is well quasi-ordered, the latter sequence is good so there exist two positive integers i, j such that $i < j$ and, for every $\ell = 1, \ldots, m$, $x_\ell^i \leqslant x_\ell^j$. The previous condition and the monotonicity of $\leqslant$ entails $w_i \leqslant w_j$. The latter contradicts that γ is bad. This proves that L is well quasi-ordered by $\leqslant$.

If the grammar G contains either unitary productions or ε-productions, the proof is almost the same. One has only to consider minimal bad sequences, assuming as a parameter the minimal length of a derivation of a word, instead of its length. $\square$

The corollary below immediately follows from condition (i) of Lemma 1 and Proposition 1.

Corollary 1. *Let $G = (V, A, P)$ be a context-free grammar where $V = \{X_1, X_2, \ldots, X_n\}$. If $\leqslant$ is a monotone quasi-order compatible with G, then L_i is well quasi-ordered by $\leqslant$ for every $i = 1, \ldots, n$.*

The following notion is a natural extension of that of division order in the free monoid.

Definition 4. Let $L \subseteq A^*$ be a language and let $\leqslant$ be a quasi-order. Then $\leqslant$ is a *division order on L* if $\leqslant$ is monotone and the following condition holds:

$$u \leqslant xuy \text{ for every } u \in L,\ x,\ y \in A^* \text{ with } xuy \in L.$$

When L is the whole free monoid A^*, the above notion coincides with the standard one of division order. On the other hand there exist orderings which have the division property on some language L and not on A^*. The following theorem holds.

Theorem 3. *Let $G = (V, A, P)$ be a context-free grammar and, according to the previous notation, let $L = \bigcup_{i=1}^{n} L_i$ be the union of all languages generated by the variables of G. If $\leqslant$ is a division order on L, then $\leqslant$ is a well quasi-order on L.*

Proof. It is easily checked that $\leqslant$ is compatible with G. Indeed, let $X_i \to p$ be a production of G. Suppose $p = u_1 Y_1 \cdots u_m Y_m u_{m+1}$ with $u_i \in A^*$, for $i = 1, \ldots, m+1$ and $Y_i \in V$, for $i = 1, \ldots, m$. Let $x_i \in L(Y_i)$ for every $i = 1, \ldots, m$. Hence $u_1 x_1 \cdots u_m x_m u_{m+1} \in L$. Since $\leqslant$ is a division order on L, one has

$$x_i \leqslant (u_1 x_1 \cdots x_{i-1} u_i) x_i (u_{i+1} x_{i+1} \cdots u_m x_m u_{m+1}),$$

for every $i = 1, \ldots, m$. The result follows from Proposition 1. $\square$

Now we give a slight generalization of the notion of division order on languages.

Definition 5. Let $L \subseteq A^*$ be a language and let $\leqslant$ be a monotone quasi-order. Then $\leqslant$ is a *weak division order on L* if for any $u, x, y \in A^*$ such that $u, xuy, xy \in L$, one has $u \leqslant xuy$.

Remark 1. We observe that any division order on L is a weak division order on L but the converse is false (see Remark 2). Moreover, any weak division order on A^* is a division order.

The following proposition is a slight extension of Theorem 3.

Theorem 4. *Let $G = (V, A, P)$ be a context-free grammar and, according to the previous notation, let $L = \bigcup_{i=1}^{n} L_i$ be the union of all the languages generated by the variables of G. Suppose that $\varepsilon \in L_i$, for any $i = 1, \ldots, n$. If $\leqslant$ is a weak division order on L, then $\leqslant$ is a well quasi-order on L.*

Proof. The proof of the claim is similar to that of Theorem 3. Indeed it is easily checked that $\leqslant$ is compatible with G. Let $X_i \rightarrow p$ be a production of G. Suppose $p = u_1 Y_1 \cdots u_m Y_m u_{m+1}$ with $u_i \in A^*$, for $i = 1, \ldots, m+1$ and $Y_i \in V$, for $i = 1, \ldots, m$. Let $x_i \in L(Y_i)$ for every $i = 1, \ldots, m$. Hence $u_1 x_1 \cdots u_m x_m u_{m+1} \in L$. Moreover, since $\varepsilon \in L_i$ for any $i = 1, \ldots, n$, one has also $(u_1 x_1 \cdots x_{i-1} u_i)(u_{i+1} x_{i+1} \cdots u_m x_m u_{m+1}) \in L$ for any $i = 1, \ldots, m$. Since $\leqslant$ is a weak division order on L, one has

$$x_i \leqslant (u_1 x_1 \cdots x_{i-1} u_i) x_i (u_{i+1} x_{i+1} \cdots u_m x_m u_{m+1}),$$

for every $i = 1, \ldots, m$.

Again, by Proposition 1, one has that $\leqslant$ is wqo on L. □

An immediate consequence of Theorem 3 and Theorem 4 is the following.

Corollary 2. *Let L be a context-free language generated by a context-free grammar with only one variable. Then any division order on L is a wqo on L. Moreover, if $\varepsilon \in L$, then any weak division order on L is a wqo on L.*

4. Well quasi-orders and unitary grammars

We now prove an interesting corollary of Proposition 1 concerning unitary semi-Thue systems. Following [5], we recall that a *rewriting system*, or *semi-Thue system* on an alphabet A is a pair (A, π) where π is a binary relation on A^*. Any pair of words $(p, q) \in \pi$ is called a *production* and denoted by $p \rightarrow q$. Let us denote by $\Rightarrow_\pi$ the derivation relation of π, that is, for $u, v \in A^*$, $u \Rightarrow_\pi v$ if

$$\exists\, (p, q) \in \pi \text{ and } \exists\, h, k \in A^* \text{ such that } u = hpk, \quad v = hqk.$$

The *derivation relation* $\Rightarrow_\pi^*$ is the transitive and reflexive closure of $\Rightarrow_\pi$. One easily verifies that $\Rightarrow_\pi^*$ is a monotone quasi-order on A^*.

A semi-Thue system is called *unitary* if π is a finite set of productions of the kind

$$\varepsilon \rightarrow u, \; u \in I, \quad I \subseteq A^+.$$

Such a system, also called *unitary grammar*, is then determined by the finite set $I \subseteq A^+$. Its derivation relation and its transitive and reflexive closure are denoted by $\Rightarrow_I$ (or, simply, $\Rightarrow$) and $\Rightarrow_I^*$ (or, simply, $\Rightarrow^*$), respectively. We set $L_I^\varepsilon = \{u \in A^* \mid \varepsilon \Rightarrow^* u\}$.

Unitary grammars have been introduced in [6], where the following theorem is proved.

Theorem 5. *Let $I \subseteq A^+$ and assume that $A = \mathrm{alph}(I)$. The following conditions are equivalent*:

(i) *the derivation relation $\Rightarrow_I^*$ is a wqo on A^**;
(ii) *the set I is subword unavoidable in A^*, that is there exists a positive integer k such that any word $u \in A^*$, with $|u| \geqslant k$, contains as a factor a word of I*;
(iii) *the language L_I^ε is regular*.

For any finite set $I \subseteq A^+$, the language L_I^ε is context-free. The construction of the grammar generating L_I^ε belongs to the folklore. We report it for completeness.

Definition 6. Let I be a finite subset of A^+. Let $G_I = (V, A, P)$ be the context-free grammar where $V = \{X\}$, $A = \mathrm{alph}(I)$ and P is the set of productions defined as
– $X \longrightarrow \varepsilon$,
– for every $u = a_1 \cdots a_n \in I$, where $a_i \in A$, $1 \leqslant i \leqslant n$,

$$X \longrightarrow Xa_1Xa_2X \cdots Xa_nX.$$

Lemma 2. *Let I be a finite subset of* A^+. *Then* $L(G_I) = L(X) = L_I^\varepsilon$.

Let I be a finite subset of A^+. Then we denote by $\vdash_I$ the binary relation of A^* defined as: for every $u, v \in A^*$, $u \vdash_I v$ if

$$u = u_1u_2 \cdots u_{n+1},$$

$$v = u_1a_1u_2a_2 \cdots u_na_nu_{n+1},$$

with $u_i \in A^*$, $a_i \in A$, and $a_1 \cdots a_n \in I$.

The relation $\vdash_I^*$ is the transitive and reflexive closure of $\vdash_I$. One easily verifies that $\vdash_I^*$ is a monotone quasi-order on A^*. Moreover $L_{\vdash_I}^\varepsilon$ denotes the set of all words derived from the empty word by applying $\vdash_I^*$, that is

$$L_{\vdash_I}^\varepsilon = \{u \in A^* \mid \varepsilon \vdash_I^* u\}.$$

The relation $\vdash_I^*$ has been considered in [8] where the following extension of Theorem 5 has been proved.

Theorem 6. *Let* $I \subseteq A^+$ *and assume that* $A = \mathrm{alph}(I)$. *The following conditions are equivalent*:
(i) *the derivation relation* $\vdash_I^*$ *is a wqo on* A^*;
(ii) *the set I is subsequence unavoidable in* A^*, *that is there exists a positive integer k such that any word* $u \in A^*$, *with* $|u| \geqslant k$, *contains as a subsequence a word of I*;
(iii) *the language* $L_{\vdash_I}^\varepsilon$ *is regular*.

Generally $\Rightarrow_I^*$ is not a wqo on L_I^ε. In fact let $A = \{a, b, c\}$, $I = \{ab, c\}$, and consider the sequence $\sigma = \{acb, aacbb, aaacbbb, \ldots, a^ncb^n, \ldots\}$. It is easy to see that the elements of σ are pairwise incomparable with respect to $\Rightarrow_I^*$, so that σ is bad. We observe that σ is not bad with respect to $\vdash^*{}_I$. Indeed for any n, m, $n \leqslant m$, one has $a^ncb^n \vdash_I^* a^mcb^m$.

Lemma 3. *Let* $x, y \in A^*$ *such that* $xy \in L_{\vdash_I}^\varepsilon$. *Then, for any* $u \in A^*$, $u \vdash_I^* xuy$.

Proof. Since $xy \in L_{\vdash_I}^\varepsilon$, one has $\varepsilon \vdash_I^n xy$ with $n \geqslant 0$. We proceed by induction on n. The basis of the induction is trivially checked. Suppose $\varepsilon \vdash_I^n xy$ with $n \geqslant 1$ so that $\varepsilon \vdash_I^{n-1} w \vdash_I xy$. Hence $w = w_1 \cdots w_{k+1}$ and $xy = w_1a_1 \cdots w_ka_kw_{k+1}$ with $a_1 \cdots a_k \in I$ and $w_i \in A^*$, for any $i = 1, \ldots, k+1$. Then $x = w_1a_1 \cdots a_{i-1}w_i'$ and $y = w_i''a_{i+1} \cdots w_{k+1}$ where $w_i = w_i'w_i''$. Now let $x' = w_1 \cdots w_i'$ and $y' = w_i'' \cdots w_{k+1}$. Hence $x'y' = w$ so, by the induction hypothesis, one has $u \vdash_I^* x'uy'$ which yields $u \vdash_I^* x'uy' = (w_1 \cdots w_i')u(w_i'' \cdots w_{k+1}) \vdash_I (w_1a_1 \cdots a_{i-1}w_i')u(w_i''a_i \cdots w_{k+1}) = xuy$. The claim is thus proved. $\square$

The following proposition immediately follows from Lemma 3.

Proposition 2. *Let $I \subseteq A^+$. Then $\vdash_I^*$ is a weak division order on L_I^ε and $L_{\vdash_I}^\varepsilon$.*

Remark 2. We observe that, in general, $\vdash^*{}_I$ is not a division order on L_I^ε. Indeed, let $A = \{a, b\}$ and let $I = \{ab, babb\}$. Set $u = ab$ and $babb = xuy$ with $x = y = b$. Then it is easily checked that $u, xuy \in L_I^\varepsilon$ but $u \not\vdash^*{}_I xuy$.

The following theorem holds.

Theorem 7. *Let I be a finite set of words. Then $\vdash^*{}_I$ is wqo on L_I^ε.*

Proof. By the latter proposition, one has that $\vdash^*{}_I$ is a weak division order on L_I^ε. Now the claim follows from Lemma 2 and Corollary 2. $\square$

Finally we consider another application of Corollary 2. For this purpose, we find it convenient to introduce some notions. A *tuple t* is a finite sequence $(t_1, \ldots, t_n)$ of words of A^+ where $n \geqslant 1$. Let T be a finite and non-empty set of tuples. Then we denote by $\leqslant_T$ the reflexive and transitive closure of the binary relation defined as

$$\{ (u, v) \in A^* \times A^* \ | \ \exists\, t = (t_1, \ldots, t_n) \in T \ |$$

$$v = u_1 t_1 u_2 t_2 \cdots u_n t_n u_{n+1}, \quad u = u_1 u_2 \cdots u_n u_{n+1}, \quad u_i \in A^*, \quad i = 1, \ldots, n+1\}.$$

The relation $\leqslant_T$ has been introduced by Haussler in [8] and it is easily checked that it generalizes both relations $\vdash_I^*$ and $\Rightarrow_I^*$.

Now we adopt the following notation. Let I be a subset of A^+. Then $\overline{I}$ denotes the following set of tuples of words

$$\overline{I} \ = \ \{(u, v) \ | \ u, v \in A^+, uv \in I\} \ \cup \ I.$$

Lemma 4. *Let $x, y \in A^*$ such that $xy \in L_I^\varepsilon$. Then, for any $u \in A^*$, one has $u \leqslant_{\overline{I}} xuy$.*

Proof. Since $xy \in L_I^\varepsilon$, one has $\varepsilon \Rightarrow_I^n xy$, $n \geqslant 0$. We proceed by induction on n. The basis of the induction is trivially checked. Let us prove the induction step. Suppose $\varepsilon \Rightarrow_I^n xy$, $n \geqslant 1$ so that $\varepsilon \Rightarrow_I^{n-1} U \Rightarrow_I xy$. Then we have the following cases:

1. $xy = (x'wx'')y$, $U = x'x''y$, where $x', x'' \in A^*$, and $w \in I$. By the induction hypothesis, one has $u \leqslant_{\overline{I}} x'x''uy$. By the definition of $\leqslant_{\overline{I}}$, one has $x'x''uy \leqslant_{\overline{I}} x'wx''uy = xuy$. Therefore $u \leqslant_{\overline{I}} xuy$.

2. $xy = x(y'wy'')$, $U = xy'y''$, where $y', y'' \in A^*$, $w \in I$. One proceeds as in (1).

3. $xy = x'wy'$, $U = x'y'$, where $x', y' \in A^*$, $w \in I$ and $x = x'w_1$, $y = w_2 y'$, $w = w_1 w_2$. We can suppose $w_1, w_2 \neq \varepsilon$, otherwise we are in case 1 or 2. By the induction hypothesis one has $u \leqslant_{\overline{I}} x'uy'$. Again, by the definition of $\leqslant_{\overline{I}}$, one has $x'uy' \leqslant_{\overline{I}} x'w_1 u w_2 y' = xuy$ which implies the result.

The proof of the claim is thus complete. $\square$

An immediate consequence of the latter lemma is the following.

Proposition 3. *The relation* $\leqslant_{\overline{I}}$ *is a weak division order on* L_I^{ε}.

Corollary 3. *The relation* $\leqslant_{\overline{I}}$ *is a wqo on* L_I^{ε}.

Proof. By the latter proposition, one has that $\leqslant_{\overline{I}}$ is a weak division order on L_I^{ε}. Now the claim follows from Lemma 2 and Corollary 2. □

5. A counterexample

In the previous section we proved that for any subset I of A^+ the relation $\vdash_I^*$ is a weak division order on L_I^{ε}. From this we derived that $\vdash_I^*$ is a wqo on L_I^{ε}. Therefore it is natural to ask whether $\vdash_I^*$ is a wqo on $L_{\vdash_I}^{\varepsilon}$ or not. The answer is negative. In fact, we now exhibit a set I such that the quasi-order $\vdash_I^*$ is not a wqo on $L_{\vdash_I}^{\varepsilon}$. For this purpose, let $A = \{a, b, c, d\}$ be a four-letter alphabet and let $\bar{A} = \{\bar{a}, \bar{b}, \bar{c}, \bar{d}\}$ be a disjoint copy of A. Let $\tilde{A} = A \cup \bar{A}$ and let $I = \{a\bar{a}, b\bar{b}, c\bar{c}, d\bar{d}\}$.

Now consider the sequence $\{S_n\}_{n \geqslant 1}$ of words of $\tilde{A}^*$ defined as: for every $n \geqslant 1$,

$$S_n = adb\bar{b}c\bar{c}\bar{a}(a\bar{d}dc\bar{c}c\bar{c}\bar{a})^n a\bar{d}b\bar{b}\bar{a}.$$

The following result holds.

Proposition 4. *The sequence* $\{S_n\}_{n \geqslant 1}$ *is bad with respect to* $\vdash_I^*$. *Moreover, the elements of* $\{S_n\}_{n \geqslant 1}$ *belong to* $L_{\vdash_I}^{\varepsilon}$ *and so* $\vdash_I^*$ *is not a wqo on* $L_{\vdash_I}^{\varepsilon}$.

Remark 3. We observe that one can easily prove that $\vdash_I^*$ is a division order on $L_{\vdash_I}^{\varepsilon}$. Therefore, if one drops the hypothesis on the structure of L, Theorem 3 does not hold any more. On the other hand the language $L_{\vdash_I}^{\varepsilon}$ is not context-free.

In order to prove Proposition 4, we need some preliminary definitions and lemmas.

Lemma 5. *Let* $u \in L_{\vdash_I}^{\varepsilon}$. *For every* $p \in Pref(u)$ *and* $x \in A$, $|p|_{\bar{x}} \leqslant |p|_x$.

Proof. $u \in L_{\vdash_I}^{\varepsilon}$ implies $\varepsilon \vdash_I^k u$, for some $k \geqslant 0$. By induction on k, one easily derives the assertion. □

The following definitions will be used later.

Definition 7. Let $u = a_1 \cdots a_n$ and $v = b_1 \cdots b_m$ be two words over $\tilde{A}$ with $n \leqslant m$. An *embedding* of u in v is a map $f : [n] \longrightarrow [m]$ such that f is increasing and, for every $i = 1, \ldots, n$, $a_i = b_{f(i)}$.

Definition 8. Let $u, v \in \tilde{A}^*$ and let *f* be an embedding of *u* in *v*. Let $v = b_1 \cdots b_m$. Then $\langle v - u \rangle_f$ is the subsequence of *v* defined as

$$\langle v - u \rangle_f = b_{i_1} \cdots b_{i_\ell} \ \text{ where, for every } k = 1, \ldots, \ell,$$

$$i_k \notin \mathrm{Im}(f).$$

The word $\langle v - u \rangle_f$ is called the *difference of v and u with respect to f.*

It is useful to remark that $\langle v - u \rangle_f$ is obtained from *v* by deleting, one by one, all the letters of *u* according to *f*.

Example 1. Let $u = a\bar{a}$ and $v = ab\bar{a}\bar{b}a\bar{a}$. Let *f* and *g* be two embeddings of *u* in *v* defined respectively as: $f(1) = 1,\ f(2) = 3,$ and $g(1) = 5,\ g(2) = 6$. Then we have $\langle v - u \rangle_f = b\bar{b}a\bar{a}$ and $\langle v - u \rangle_g = ab\bar{a}\bar{b}$.

Remark 4. A word *u* is a subsequence of *v* if and only if there exists an embedding of *u* in *v*.

Remark 5. An embedding *f* of *u* in *v* is uniquely determined by two factorizations of *u* and *v* of the form

$$u = a_1 a_2 \cdots a_n, \qquad v = v_1 a_1 v_2 a_2 \cdots v_n a_n v_{n+1}$$

with $a_i \in \tilde{A},\ v_i \in \tilde{A}^*$.

In the sequel, according to the latter remark, $\langle v - u \rangle_f$ may be written as

$$\langle v - u \rangle_f = v_1 v_2 \cdots v_n v_{n+1}.$$

Lemma 6. *Let $u, v \in L^{\varepsilon}_{\vdash_I}$ such that $u \vdash_I^* v$. Then there exists an embedding f of u in v such that*

$$\langle v - u \rangle_f \ \in \ L^{\varepsilon}_{\vdash_I}.$$

Proof. The proof is by induction. By hypothesis there exists $k \geqslant 0$ such that $u \vdash_I^k v$. If $k = 0$, then $u = v$ so $\langle v - u \rangle_f = \varepsilon \in L^{\varepsilon}_{\vdash_I}$. Suppose $k = 1$. Thus $u = u_1 u_2 u_3$ and $v = u_1 x u_2 \bar{x} u_3$ where $x \in A$ and $u_1 u_2 u_3 \in L^{\varepsilon}_{\vdash_I}$. Hence $\langle v - u \rangle_f = x\bar{x} \in L^{\varepsilon}_{\vdash_I}$. The basis of the induction is proved.

Let us prove the induction step. Suppose $u \vdash_I^{k+1} v$ with $k \geqslant 1$. Then there exists $w \in L^{\varepsilon}_{\vdash_I}$ such that $u \vdash_I^k w$ and $w \vdash_I v$. By the induction hypothesis, there exists an embedding *f* of *u* in *w* such that $\langle w - u \rangle_f \in L^{\varepsilon}_{\vdash_I}$. Suppose $u = a_1 \cdots a_n$ and $w = u_1 a_1 u_2 a_2 \cdots u_i a_i \cdots u_n a_n u_{n+1}$ with $a_i \in \tilde{A}$, $u_i \in \tilde{A}^*$. Hence $\langle w - u \rangle_f = u_1 u_2 \cdots u_{n+1} \in L^{\varepsilon}_{\vdash_I}$. Since $w \vdash_I v$, suppose that

$$v = u_1 a_1 u_2 a_2 \cdots u_i x \cdots u_j \bar{x} \cdots u_n a_n u_{n+1},$$

with $x \in A$ (the other cases determined by different positions of x and $\bar{x}$ are treated similarly). From the latter condition, one easily sees that f may be extended to an embedding g of u in v such that

$$\langle v - u \rangle_g = u_1 u_2 \cdots u_i x \cdots u_j \bar{x} \cdots u_n u_{n+1}.$$

Since $\langle w - u \rangle_f \in L^{\varepsilon}_{\vdash_I}$ and $\langle w - u \rangle_f \vdash_I \langle v - u \rangle_g$, one has $\langle v - u \rangle_g \in L^{\varepsilon}_{\vdash_I}$. $\square$

Lemma 7. *For every $m, n \geqslant 1$ one has*:
(i) $S_n \in L^{\varepsilon}_{\vdash_I}$;
(ii) $S_n \in Fact(S_m)$ *if and only if* $n = m$;
(iii) *Suppose* $n \leqslant m$. *Let* $Q = adb\bar{b}c\bar{c}\bar{a}(a\bar{d}dc\bar{c}c\bar{c}\bar{a})^n a\bar{d}$. *Then* $Q \in Pref(S_n) \cap Pref(S_m)$.

Proof. By induction on n, condition (i) is easily proved. Conditions (ii) and (iii) immediately follow from the structure of words of $\{S_n\}_{n \geqslant 1}$. $\square$

Lemma 8. *Let n, m be positive integers such that $n \leqslant m$. If $S_n \vdash_I^* S_m$ then $S_n = S_m$.*

Proof. Let $n \leqslant m$ be positive integers. Then

$$S_n = adb\bar{b}c\bar{c}\bar{a}(a\bar{d}dc\bar{c}c\bar{c}\bar{a})^n a\bar{d}b\bar{b}\bar{a} \text{ and}$$

$$S_m = adb\bar{b}c\bar{c}\bar{a}(a\bar{d}dc\bar{c}c\bar{c}\bar{a})^n (a\bar{d}dc\bar{c}c\bar{c}\bar{a})^k a\bar{d}b\bar{b}\bar{a}, \text{ with } k \geqslant 0.$$

By Lemma 6, the hypothesis $S_n \vdash_I^* S_m$ implies there exists an embedding f of S_n in S_m such that $\langle S_m - S_n \rangle_f \in L^{\varepsilon}_{\vdash_I}$.

We now prove the following claim.

Claim. *The following conditions hold*:

(1) *For all $i = 1, \ldots, 9 + 8n$, $f(i) = i$. In particular, by condition* (iii) *of Lemma 7, f is the identity on the common prefix $Q = adb\bar{b}c\bar{c}\bar{a}(a\bar{d}dc\bar{c}c\bar{c}\bar{a})^n a\bar{d}$ of S_n and S_m.*
(2) $f(|S_n| - i) = |S_m| - i, \quad for\ i = 0, 1, 2.$

Proof of the Claim. First we observe that, for all $n \geqslant 1$, $b\bar{b}$ occurs exactly twice as a factor of S_n. This immediately entails condition (2) and $f(i) = i$ for all $i = 1, \ldots, 4$.

The proof of condition (1) is divided into the following two steps.

Step 1: Let i be a positive integer such that $i \leqslant 9 + 8n$. If $a_i \in \{a, \bar{a}, d, \bar{d}\}$, then $f(i) = i$.

We first observe that, for all i such that $4 \leqslant i \leqslant 9 + 8n$, one has:
– If $a_i = d$ (resp. $a_i = \bar{d}$) then $i = 10 + 8\ell$ (resp. $i = 9 + 8\ell$), with $\ell \geqslant 0$;
– If $a_i = a$ (resp. $a_i = \bar{a}$) then $i = 8(\ell + 1)$ (resp. $i = 8(\ell + 1) - 1$), with $\ell \geqslant 0$.
Now we prove Step 1 by induction on $\ell \geqslant 0$. One easily checks that $f(2) = 2$ yields $f(9) = 9$. Indeed, if $f(9) > 9$ then $\langle S_m - S_n \rangle_f = v'v''$, with $v', v'' \in \tilde{A}^*$ and $|v'|_{\bar{d}} = 1 > |v'|_d = 0$. By Lemma 5, $\langle S_m - S_n \rangle_f \notin L^{\varepsilon}_{\vdash_I}$ which contradicts the choice of f. Hence $f(9) = 9$. This entails $f(7) = 7$ and $f(8) = 8$.

By using a similar argument, conditions $f(10) = 10$ and $f(15) = 15$ follow from $f(8) = 8$. The basis of the induction is proved.

Let us prove the induction step. Let $i = 10 + 8(\ell - 1)$. Then $a_i = d$ and, by induction hypothesis, $f(i) = i$. This yields $f(9 + 8\ell) = 9 + 8\ell$. Indeed, otherwise, $\langle S_m - S_n\rangle_f = v'v''$, with $v',\ v'' \in \tilde{A}^*$ and $|v'|_{\bar{d}} = 1 > |v'|_d = 0$. As before, $\langle S_m - S_n\rangle_f \notin L^{\varepsilon}_{\vdash_I}$ which contradicts the choice of f. Hence $f(9+8\ell) = 9+8\ell$ which entails $f(8(\ell+1)) = \bar{8}(\ell+1)$ and $f(8(\ell+1)-1) = 8(\ell+1)-1$. By using a similar argument from the latter condition one derives $f(10 + 8\ell) = 10 + 8\ell$. This proves Step 1.

Step 2: Let i be a positive integer such that $i \leqslant 9 + 8n$. If $a_i \in \{c,\ \bar{c}\}$, then $f(i) = i$.

First we observe that every occurrence of $c\bar{c}$ in S_n is a factor of an occurrence of $db\bar{b}c\bar{c}\bar{a}$ or $dc\bar{c}c\bar{c}\bar{a}$. Let us consider the second case (the first is similarly treated). Set $dc\bar{c}c\bar{c}\bar{a} = a_i \cdots a_{i+5}$ with $i \geqslant 1$. By Step 1, $f(i) = i$ and $f(i+5) = i+5$ which immediately entails $f(i+\ell) = i+\ell$, for $\ell = 1, \ldots, 4$. This proves Step 2.

Finally, Condition (1) follows from Steps 1 and 2. $\diamond$

Suppose now $k > 0$. Then the previous claim implies

$$\langle S_m - S_n\rangle_f = dc\bar{c}c\bar{c}\bar{a}(a\bar{\bar{d}}dc\bar{c}c\bar{c}\bar{a})^{k-1}a\bar{d}.$$

Let $p = dc\bar{c}c\bar{c}\bar{a}$. Since $p \in Pref(\langle S_m - S_n\rangle_f)$ and $|p|_{\bar{a}} > |p|_a$, Lemma 5 implies $\langle S_m - S_n\rangle_f \notin L^{\varepsilon}_{\vdash_I}$. Hence the case $n < m$ is not possible. This proves the Lemma. $\square$

Proof of Proposition 4. We prove the claim by contradiction. Thus there exist $n, m \geqslant 1$ such that $n < m$ and $S_n \vdash_I^* S_m$. By Lemma 8, $S_n = S_m$. Hence, by condition (ii) of Lemma 7, $n = m$ which is a contradiction. This proves that the sequence $\{S_n\}_{n\geqslant 1}$ is bad.

References

[1] J. Berstel, Transductions and Context-Free Languages, Teubner, Stuttgart, 1979.
[2] D.P. Bovet, S. Varricchio, On the regularity of languages on a binary alphabet generated by copying systems, Inform. Process. Lett. 44 (1992) 119–123.
[3] A. de Luca, S. Varricchio, Some regularity conditions based on well quasi-orders, Lecture Notes in Computer Science, vol. 583, Springer, Berlin, 1992, pp. 356–371.
[4] A. de Luca, S. Varricchio, Well quasi-orders and regular languages, Acta Informatica 31 (1994) 539–557.
[5] A. de Luca, S. Varricchio, Finiteness and regularity in semigroups and formal languages, EATCS Monographs on Theoretical Computer Science, Springer, Berlin, 1999.
[6] A. Ehrenfeucht, D. Haussler, G. Rozenberg, On regularity of context-free languages, Theoret. Comput. Sci. 27 (1983) 311–332.
[7] T. Harju, L. Ilie, On well quasi orders of words and the confluence property, Theoret. Comput. Sci. 200 (1998) 205–224.
[8] D. Haussler, Another generalization of Higman's well quasi-order result on Σ^*, Discrete Math. 57 (1985) 237–243.
[9] G.H. Higman, Ordering by divisibility in abstract algebras, Proc. London Math. Soc. 3 (1952) 326–336.
[10] L. Ilie, A. Salomaa, On well quasi orders of free monoids, Theoret. Comput. Sci. 204 (1998) 131–152.
[11] B. Intrigila, S. Varricchio, On the generalization of Higman and Kruskal's theorems to regular languages and rational trees, Acta Informatica 36 (2000) 817–835.
[12] J. Kruskal, The theory of well-quasi-ordering: a frequently discovered concept, J. Combin. Theory, Ser. A 13 (1972) 297–305.