Well quasi orders on languages ⋆

Flavio D'Alessandro[1] and Stefano Varricchio[2]

[1] Dipartimento di Matematica, Università di Roma "La Sapienza" Piazzale Aldo Moro 2, 00185 Roma, Italy `dalessan@mat.uniroma1.it`

[2] Dipartimento di Matematica, Università di Roma "Tor Vergata", via della Ricerca Scientifica, 00133 Roma, Italy. `varricch@mat.uniroma2.it`

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.

1 Introduction

A *quasi-order* on a set S is called a *well quasi-order* (*wqo*) if every nonempty 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 $\leq$. If there exists a generating set of S well quasi-ordered by $\leq$, then S will also be so.* From this one derives that the *subsequence ordering* in free monoids is a wqo.

In [12] Kruskal extends Higman's result, proving that some 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].

⋆ This work was partially supported by MIUR project `Linguaggi formali e automi: teoria e applicazioni`.

In this paper we give a new generalization of Higman's theorem. First of all we give the notion of *division order* on a language L: a quasi order $\leq$ 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 $u \leq v$. 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.

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

$$\{\epsilon \rightarrow u,\ u \in I\}.$$

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 $L_I^\epsilon = \{w \in A^* \mid \epsilon \Rightarrow_I^* w\}$, $L_{\vdash_I}^\epsilon = \{w \in A^* \mid \epsilon \vdash_I^* w\}$ 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 $L_{\vdash_I}^\epsilon$.
- For any finite set I the relation $\vdash_I^*$ is a wqo on L_I^ϵ.

Finally we observe that for any finite set I, the relation $\vdash_I^*$ is a division order on the language $L_{\vdash_I}^\epsilon$. Therefore, our main theorem does not hold in general on an arbitrary language. On the other hand the language $L_{\vdash_I}^\epsilon$ is not context-free.

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 word $w \in A^*$ can be written uniquely as a sequence of letters as $w = a_1 a_2 \cdots a_n$, with $a_i \in A$, $1 \leq i \leq n$, $n > 0$. 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. $Suf(w)$, $F(w)$).

A subset L of A^* is called a *language.* If L is a language of A^*, then alph(L) is the smallest subset B of A such that $L \subseteq B^*$. A binary relation $\leq$ on a set S is a *quasi-order* (qo) if $\leq$ is reflexive and transitive. Moreover, if $\leq$ is symmetric, then $\leq$ is an equivalence relation. The meet $\leq \cap \leq^{-1}$ 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 $\leq$ if, for every $x \in X$, $x \leq s$ implies $x \sim s$. For $s, t \in S$ if $s \leq t$ 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 $\leq$ if the following condition is satisfied:

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

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 $\leq$, if $\leq$ 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 $\leq$. The following conditions are equivalent:*

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

Let $\sigma = \{s_i\}_{i \geq 1}$ 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 $i < j$, $s_i \not\leq s_j$. It is worth noting that, by condition (iv) above, a useful technique to prove that $\leq$ 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], Ch. 6).

Lemma 1. *The following conditions hold:*

1) Every subset of a wqo set is wqo;

2) If S and T are wqo by $\leq_S$ and $\leq_T$ respectively, then $S \times T$ is wqo by $\leq_S \otimes \leq_T$.

Let us now suppose that the set S is a semigroup.

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

$$x_1 \leq x_2 \ \textit{implies}\ x_1 y \ \leq \ x_2 y \ \ (y x_1 \ \leq \ y x_2).$$

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

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

$$s \ \leq \ xsy.$$

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 $\leq$. If there exists a generating set of S well quasi-ordered by $\leq$ 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 images of f.

3 Main result

We now 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 $\Rightarrow$. 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 $\leq$ be a quasi-order on A^*. Then $\leq$ is said to be* compatible *with G if the following condition holds:*

for every production of G of the kind $X_i \ \longrightarrow \ u_1 Y_1 u_2 Y_2 \cdots u_m Y_m u_{m+1}$, where, $u_k \in A^*$, for $k = 1, \ldots, m+1$, and $Y_k \in V$, $k = 1, \ldots, m$, one has:

$$x_k \ \leq \ u_1 x_1 u_2 x_2 \cdots u_m x_m u_{m+1},$$

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 $\leq$ is a monotone quasi-order compatible with G, then $\leq$ 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 does not contain neither unitary productions nor ϵ-productions. By contradiction, deny the claim of the proposition. Hence there exists a bad sequence in L. 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. Then select a word $v_2 \in L$ such that $v_1,\ v_2$ (in that order) are the first two terms of a bad sequence in L and $|v_2|$ is as small as possible. Then select a word $v_3 \in L$ such that $v_1,\ v_2,\ v_3$ (in that order) are the first three terms of a bad sequence in L and $|v_3|$ is as small as possible. Assuming the Axiom of Choice, this process yields a bad sequence $\gamma = \{v_i\}_{i\geq 1}$ in L. This sequence is minimal in the following sense: let $\alpha = \{z_i\}_{i\geq 1}$ 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 $|v_{k+1}| \leq |z_{k+1}|$.

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

$$\forall\, \ell \geq 1,\ \ X_k \ \Rightarrow\ p\ \Rightarrow^*\ v_{i_\ell}, \qquad (1)$$

where $X_k\ \rightarrow\ p$ 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 $\ell \geq 1$, $w_\ell = v_{i_\ell}$. Hence, by (1), for every $\ell \geq 1$, one has

$$w_\ell\ =\ u_1x_1^\ell u_2x_2^\ell \cdots u_mx_m^\ell u_{m+1}, \quad \text{with}$$

$$x_1^\ell \in L(Y_1),\ \ x_2^\ell \in L(Y_2),\ \ \ldots,\ \ x_m^\ell \in L(Y_m).$$

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

Claim. *For every $j = 1, \ldots, m$, F_j is wqo by $\leq$.*

Proof of the Claim: By contradiction, let j be a positive integer with $1 \leq j \leq m$ such that F_j is not wqo by $\leq$. Let $\tau = \{t_i\}_{i\geq 1}$ be a bad sequence in F_j.

We first observe that, for all $i \geq 1$, there exists a positive integer $g(i)$ such that $t_i = x_j^{g(i)}$. Hence, we can consider a subsequence of τ, say $\{y_i\}_{i\geq 1}$, such that, for every $i \geq 1$, $y_i = x_j^{g(i)}$ with $g(i) \geq g(1)$.

Consider now the sequence

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

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

$$x_j^{g(k)}\ \leq\ u_1x_1^{g(k)}u_2\cdots u_mx_m^{g(k)}u_{m+1}\ =\ v_{g(k)}.$$

Hence $v_h \leq v_{g(k)}$. Since $h < g(1) \leq g(k)$ the latter contradicts that γ is bad. Hence $v_h \not\leq y_k$.

Now we observe that $|y_1| < |v_{g(1)}|$ contradicts that γ is minimal. Hence, no bad sequence in F_j exists so F_j is wqo by $\leq$. $\diamond$

Let $\mathcal{F} = F_1 \times F_2 \times \cdots \times F_j \times \cdots \times F_m$. By condition (2) of Lemma 1 and the claim above, one has the set $\mathcal{F}$ is wqo by the canonical extension of $\leq$ on $\mathcal{F}$. Consider now the sequence of $\mathcal{F}$ defined as

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

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

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.

□

The corollary below immediately follows from condition (1) 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 $\leq$ is a monotone quasi-order compatible with G, then L_i is wqo by $\leq$ 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 $\leq$ be a quasi-order. Then $\leq$ is a* division order *on L if $\leq$ is monotone and the following condition holds:*

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

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 G. The following theorem holds.

Theorem 3. *If $\leq$ is a division order on L, then $\leq$ is a well quasi-order on L.*

Proof. It is easily checked that $\leq$ is compatible with G. Indeed, let $X_i \ \rightarrow\ p$ be a production of G. Suppose $p = u_1Y_1 \cdots u_mY_mu_{m+1}$ 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 $u_1x_1 \cdots u_mx_mu_{m+1} \in L$. Since $\leq$ is a division order on L, one has

$$x_i \ \leq\ (u_1x_1 \cdots x_{i-1}u_i)x_i(u_{i+1}x_{i+1} \cdots u_mx_mu_{m+1}),$$

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

Then the result follows from Proposition 1. □

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 $(A,\ \pi)$ where π is a binary relation on A^*. Any pair of words $(p, q) \in \pi$ is called a *production* and denoted by $p \to q$. Let us denote by $\Rightarrow_\pi$ the derivation relation of π, that is, for $u, v \in A^*$, $u \ \Rightarrow_\pi \ v$ if and only if

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

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

$$\epsilon \to u,\ u \in I,\ \ I \subseteq A^*.$$

Such a system is then determined by the finite set $I \subseteq A^*$. The derivation relation of it and its regular closure are denoted by $\Rightarrow_I$ (or, simply, $\Rightarrow$) and $\Rightarrow_I^*$ (or, simply, $\Rightarrow^*$), respectively. We set $L_I^\epsilon = \{u \in A^* \ \mid \ \epsilon \ \Rightarrow^* \ u\}$.

The following Lemma states that a unitary semi-Thue system may be simulated by a suitable context-free grammar and it belongs to the folklore.

Definition 5. *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:*

– $X \ \longrightarrow \ \epsilon,$

– *for every $u = a_1 \cdots a_n \in I$, where $a_i \in A,\ 1 \leq i \leq n$,*

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

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

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 $\vdash_\pi^*$ is a monotone quasi-order on A^*. Moreover $L_{\vdash_I}^\epsilon$ denotes the set of all words derived from the empty word by applying $\vdash_I^*$, that is

$$L^{\epsilon}_{\vdash_I} = \{u \in A^* \mid \epsilon \vdash^*_I u\}.$$

Generally $\Rightarrow^*_I$ is not a wqo on L^{ϵ}_I. In fact let $A = \{a, b, c\}$ and $I = \{ab, c\}$. Then the sequence $acb, aacbb, aaacbbb, \ldots, a^n cb^n \ldots$ is a bad sequence with respect to $\Rightarrow^*_I$. The following theorem holds.

Theorem 4. *Let I be a finite set of words. Then $\vdash^*{}_I$ is wqo on L^{ϵ}_I.*

Proof. First we prove that $\vdash^*{}_I$ is compatible with the grammar G_I. According to the definition of G_I and by Lemma 2, the task amounts to show that, for every $v = a_1 \cdots a_m \in I$ and, for every $u_1, \ldots, u_{m+1} \in L^{\epsilon}_I$, one has, for $i = 1, \ldots m+1$,

$$u_i \vdash^*_I u_1 a_1 u_2 a_2 \cdots u_m a_m u_{m+1}.$$

Since $\Rightarrow^*_I$ is monotone and $u_i \in L^{\epsilon}_I$, for every $i = 1, \ldots, m+1$, one has $u_i \Rightarrow^*_I u_1 \cdots u_{m+1}$. So $u_i \vdash^*_I u_1 \cdots u_{m+1} \vdash^*_I u_1 a_1 \cdots a_m u_{m+1}$. Finally, the claim follows from Proposition 1 and Lemma 2.

We prove that, for a suitable finite set I of words over a finite alphabet, the quasi-order $\vdash^*_I$ is not a wqo on $L^{\epsilon}_{\vdash_I}$.

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 $\{S_n\}_{n \geq 1}$ of words of $\tilde{A}^*$ defined as: for every $n \geq 1$,

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

The following result holds.

Proposition 2. *$\{S_n\}_{n \geq 1}$ is a bad sequence in $\tilde{A}^*$ with respect to $\vdash^*_I$. Hence $\vdash^*_I$ is not wqo on $\tilde{A}^*$. In particular, $\vdash^*_I$ is not wqo on $L^{\epsilon}_{\vdash_I}$.*

Remark 1. We observe that one can easily prove that $\vdash^*_I$ is a division order on $L^{\epsilon}_{\vdash_I}$. Therefore, if one drops the hypothesis on the structure of L, Theorem 3 does not hold any more. On the other hands the language $L^{\epsilon}_{\vdash_I}$ is not context-free.

In order to prove Proposition 2, we need some preliminary definitions and lemmas.

Lemma 3. *Let $u \in L^{\epsilon}_{\vdash_I}$. For every $p \in Pref(u)$ and $x \in A$, $|p|_{\bar{x}} \leq |p|_x$.*

Proof. $u \in L^{\epsilon}_{\vdash_I}$ implies $\epsilon \vdash^k_I u$, for some $k \geq 0$. By induction on k, one easily derives the assertion. □

The following definitions will be used later.

Definition 6. *Let $u = a_1 \cdots a_n$ and $v = b_1 \cdots b_m$ be two words over $\tilde{A}$ with $n \leq m$. 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 7. *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 subword 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 2. A word u is a subsequence of v if and only if there exists an embedding of u in v.

Remark 3. An embedding f of u in v is uniquely determined by two factorizations of u and v of the form

$$u = u_1 u_2 \cdots u_n, \qquad v = v_1 u_1 v_2 u_2 \cdots v_n u_n v_{n+1}$$

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

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 4. *Let $u, v \in L^\epsilon_{\vdash_I}$ such that $u \vdash_I^* v$. Then there exists an embedding f of u in v such that*

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

Proof. By induction on $k \geq 0$ such that $u \vdash_I^k v$. If $k = 0$, then $u = v$ so $\langle v - u \rangle_f = \epsilon \in L^\epsilon_{\vdash_I}$. 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 $u_1 u_2 u_3 \in L^\epsilon_{\vdash_I}$. Hence $\langle v - u \rangle_f = x\bar{x} \in L^\epsilon_{\vdash_I}$. The basis of the induction is proved.

Let us prove the induction step. Suppose $u \vdash_I^{k+1} v$ with $k \geq 1$. Then there exists $w \in L^\epsilon_{\vdash_I}$ 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 $\langle w - u \rangle_f \in L^\epsilon_{\vdash_I}$. 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 $\langle w - u \rangle_f = u_1 u_2 \cdots u_{n+1} \in L^\epsilon_{\vdash_I}$. 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 $\langle v - u \rangle_g = u_1 u_2 \cdots u_i x \cdots u_j \bar{x} \cdots u_n u_{n+1}$. □

Lemma 5. *For every $m, n \geq 1$,*

(1) $S_n \in L^{\epsilon}_{\vdash_I}$;

(2) $S_n \in F(S_m)$ if and only if $n = m$;

3) Suppose $n \leq m$. 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 (1) is easily proved. Conditions (2) and (3) immediately follow from the structure of words of $\{S_n\}_{n\geq 1}$. □

Lemma 6. *Let n, m be positive integers such that $n \leq m$. If $S_n \vdash_I^* S_m$ then $S_n = S_m$.*

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

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

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

By Lemma 4, the hypothesis $S_n \vdash_I^* S_m$ implies there exists an embedding f of S_n in S_m such that $\langle S_m - S_n\rangle_f \in L^{\epsilon}_{\vdash_I}$. We now prove the following claim.

Claim. *The following conditions hold:*

1) $\forall\, i = 1, \ldots, 9+8n$, $f(i) = i$. In particular, by condition (3) of Lemma 5, 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) $f(|S_n| - i) = |S_m| - i$, for $i = 0, 1, 2$.

Proof of the Claim: First we observe that, for all $n \geq 1$, $b\bar{b}$ 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 $i \leq 9+8n$. If $a_i \in \{a,\ \bar{a}, d,\ \bar{d}\}$, then $f(i) = i$.*
We first observe that, for all i such that $4 \leq i \leq 9 + 8n$, one has:

– If $a_i = d$ (resp. $a_i = \bar{d}$) then $i = 10 + 8\ell$ (resp. $i = 9 + 8\ell$), with $\ell \geq 0$;

– If $a_i = a$ (resp. $a_i = \bar{a}$) then $i = 8(\ell + 1)$ (resp. $i = 8(\ell + 1) - 1$), with $\ell \geq 0$.

Now we prove Step 1 by induction on $\ell \geq 0$. 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 3, $\langle S_m - S_n\rangle_f \notin L^{\epsilon}_{\vdash_I}$ 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, $\langle S_m - S_n\rangle_f \notin L^{\epsilon}_{\vdash_I}$ which contradicts the choice of f. Hence $f(9 + 8\ell) = 9 + 8\ell$ which entails $f(8\ell) = 8\ell$. This proves Step 1.

Step 2. *Let i be a positive integer such that $i \leq 9 + 8n$. 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 $i \geq 1$. 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 Step 1 and Step 2. $\diamond$

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

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

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 3 implies $\langle S_m - S_n\rangle_f \notin L^{\epsilon}_{\vdash_I}$. Hence the case $n < m$ is not possible. This proves the Lemma. □

Proof of Proposition 2: By contradiction, deny. Thus there exists $n, m \geq 1$ such that $n < m$ and $S_n \vdash_I^* S_m$. By Lemma 6, $S_n = S_m$. Hence, by condition (2) of Lemma 5, $n = m$ which is a contradiction. This proves that the sequence $\{S_n\}_{n\geq 1}$ is bad.

References

1. J. Berstel, *Transductions and Context-Free Languages.* Teubner, Stuttgart, 1979
2. D. P. Bovet and S. Varricchio, On the regularity of languages on a binary alphabet generated by copying systems. *Information Processing Letters* **44**, 119–123 (1992).
3. A. de Luca and S. Varricchio, Some regularity conditions based on well quasi-orders. *Lecture Notes in Computer Science*, Vol. 583, pp. 356–371, Springer-Verlag, Berlin, 1992.
4. A. de Luca and S. Varricchio, Well quasi-orders and regular languages. *Acta Informatica* **31**, 539–557 (1994).
5. A. de Luca and S. Varricchio, *Finiteness and regularity in semigroups and formal languages.* EATCS Monographs on Theoretical Computer Science. Springer, Berlin, 1999.
6. A. Ehrenfeucht, D. Haussler, and G. Rozenberg, On regularity of context-free languages. *Theoretical Computer Science* **27**, 311–332 (1983).
7. T. Harju and L. Ilie, On well quasi orders of words and the confluence property. *Theoretical Computer Science* **200** 205–224 (1998).

8. D. Haussler, Another generalization of Higman's well quasi-order result on Σ^*. *Discrete Mathematics* **57**, 237–243 (1985).
9. G. H. Higman, Ordering by divisibility in abstract algebras. *Proc. London Math. Soc.* **3**, 326–336 (1952).
10. L. Ilie and A. Salomaa, On well quasi orders of free monoids. *Theoretical Computer Science* **204** 131–152 (1998).
11. B. Intrigila and S.Varricchio, On the generalization of Higman and Kruskal's theorems to regular languages and rational trees. *Acta Informatica* **36**, 817–835 (2000).
12. J. Kruskal, The theory of well-quasi-ordering: a frequently discovered concept. *J. Combin. Theory, Ser. A,* **13**, 297–305 (1972).