RAIRO-Inf. Theor. Appl. **40** (2006) 407-426
DOI: 10.1051/ita:2006019

WELL QUASI-ORDERS, UNAVOIDABLE SETS, AND DERIVATION SYSTEMS *

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

Abstract. Let I be a finite set of words and $\Rightarrow_I^*$ be the derivation relation generated by the set of productions $\{\epsilon \to u \mid u \in I\}$. Let L_I^ϵ be the set of words u such that $\epsilon \Rightarrow_I^* u$. We prove that the set I is unavoidable if and only if the relation $\Rightarrow_I^*$ is a well quasi-order on the set L_I^ϵ. This result generalizes a theorem of [Ehrenfeucht *et al., Theor. Comput. Sci.* **27** (1983) 311–332]. Further generalizations are investigated.

Mathematics Subject Classification. 68Q45, 68R15.

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.

A set of words I is called *unavoidable* if there exists an integer $k > 0$ such that any word $w \in A^+$, with $A = \mathrm{alph}(I)$ and $|w| \geq k$, contains as a factor a word of I. A finite set I is called *avoidable* if it is not unavoidable.

Well quasi-orders have been widely investigated in the past. We recall the celebrated Higman and Kruskal results [10,15]. Higman gives a very general theorem on division orders in abstract algebras from which one derives that the *subsequence ordering* in free monoids is a wqo. Kruskal extends Higman's result, proving that

Keywords and phrases. Well quasi-orders, context-free languages.

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

[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

certain embeddings on finite trees are well quasi-orders. Some remarkable extensions of the Kruskal theorem are given in [12, 16].

In the last years many papers have been devoted to the applications of wqo's to formal language theory [1–8, 11].

In [7], a remarkable class of grammars, called *unitary grammars*, has been introduced in order to study the relationships between the classes of context-free and regular languages. If I is a finite set of words then we can consider the set of productions

$$\{\epsilon \to u,\ u \in I\}$$

and the derivation relation $\Rightarrow_I^*$ of the semi-Thue system associated with I. Moreover, the language generated by the unitary grammar associated with I is $L_I^\epsilon = \{w \in A^* \mid \epsilon \Rightarrow_I^* w\}$. Unavoidable sets of words are characterized in terms of the wqo property of the unitary grammars. Precisely it is proved that I is unavoidable if and only if the derivation relation $\Rightarrow_I^*$ is a wqo.

In this paper we give the following improvement of the previous result of [7]: *A finite set of words I is unavoidable if and only if the relation $\Rightarrow_I^*$ is a well quasi-order on the language L_I^ϵ.* The crucial step of our main result is the construction of a *bad sequence* of elements of L_I^ϵ, when I is avoidable. As a consequence of our theorem and of some results of [7], one obtains the equivalence of the following conditions:

- I is unavoidable;
- L_I^ϵ is regular;
- $\Rightarrow_I^*$ is a well quasi-order on L_I^ϵ.

It is worth noticing that the problems we have discussed above, may be considered with respect to other quasi-orders. In [9], Haussler investigated the relation $\vdash_I^*$ defined 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 $a_1 a_2 \cdots a_n \in I$. In particular, a characterization of the wqo property of $\vdash_I^*$ in terms of subsequence unavoidable sets of words was given in [9]. In the last part of the paper, we focus our attention on a possible extension of our main result with respect to $\vdash_I^*$.

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 let A^* be 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 non-empty word $w \in A^*$ can be written uniquely as a sequence of letters as $w = a_1 a_2 \cdots a_n$, with $a_i \in A$, $1 \le i \le n$, $n > 0$. 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. If w is the empty word, then we set $|w| = 0$ and, for any $a \in A$, $|w|_a = 0$. 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 $\mathrm{Pref}(w)$ (resp. $\mathrm{Suff}(w)$, $\mathrm{Fact}(w)$). A word u is a *subsequence* of a word v if $u = a_1 a_2 \cdots a_n$, $v = v_1 a_1 v_2 a_2 \cdots v_n a_n v_{n+1}$ 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 $L \subseteq B^*$. Moreover, $\mathrm{Pref}(L)$ denotes the set of the prefixes of all words of L. A language of A^* is called *recognizable* if it is accepted by a finite automaton or, equivalently, *via* the well known characterization of Myhill and Nerode, if it is saturated by a finite index congruence of A^*. The family of recognizable languages of A^* is denoted $Rec(A^*)$. 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). A quasi-order $\leq$ in a semigroup S is *monotone on the right (resp. on the left)* if for all $x_1, x_2, y \in S$

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

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

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 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 exist several conditions which characterize the concept of well quasi-order and that can be assumed as equivalent definitions (cf. [6]).

Theorem 2.1. *Let S be a set quasi-ordered by $\leq$. The following conditions are equivalent:*

i. *$\leq$ is a well quasi-order;*
ii. *every infinite sequence of elements of S has an infinite ascending subsequence;*
iii. *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 $s_i \leq s_j$;*
iv. *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.*

A partial order satisfying the wqo property is also called a *well partial order*. The quasi-orders considered in this paper are actually partial orders. However, according to the current terminology, we refer to them as quasi-orders.

Let $\sigma = \{s_i\}_{i\geq 1}$ be an infinite sequence of elements of S. Then σ is called *good* if it satisfies condition (iii) of Theorem 2.1 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 (iii) 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 [6], Chap. 6).

Lemma 2.2. *The following conditions hold:*

i. *every subset of a wqo set is wqo;*
ii. *if S and T are wqo by $\leq_S$ and $\leq_T$ respectively, then $S \times T$ is wqo by $\leq_S \otimes \leq_T$.*

Following [6], 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

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

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, 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 $L_I^\epsilon = \{u \in A^* \mid \epsilon \Rightarrow^* u\}$.

Unitary grammars have been introduced in [7], where the following theorem is proved.

Theorem 2.3. *Let I be a finite set of 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 unavoidable;*
iii. *the language L_I^ϵ is regular.*

3. Main result

The main result of this section will be stated in Corollary 3.23 which is an improvement of Theorem 2.3 where condition (i) is replaced by the weaker condition

that L_I^ϵ is well quasi-ordered by the relation $\Rightarrow_I^*$. In order to achieve this result we have first to prove the following non-trivial theorem.

Theorem 3.1. *Let I be a finite set of words. If I is avoidable then $\Rightarrow_I^*$ is not a wqo on the language L_I^ϵ.*

The proof of Theorem 3.1 is divided into the following four cases.

3.1. First case

We suppose that $\mathrm{Card}(I) = 1$ so that $I = \{w\}$. Set $A = \mathrm{alph}(I)$. Let us first observe that $\mathrm{Card}(A) \geq 2$. Indeed, if $A = \{a\}$, then $w = a^k$, $k \geq 1$ so that I is an unavoidable set of A^* which contradicts the assumption on the set I. Hence, w may be factorized as $w = w'ab^k$, where $a, b \in A, a \neq b, w' \in A^*$ and $k > 0$.

Now we construct the bad sequence of L_I^ϵ. For any $n > 0$, let x_n be the word defined as

$$x_n = (w'a)^{n-1}w(w'a)b^{kn}.$$

The following lemma states some useful properties of the words of the sequence $\{x_n\}$.

Lemma 3.2. *The following conditions hold:*

i. *for any $n > 0$, $x_n \in L_I^\epsilon$;*
ii. *for any $n > 0$, $|x_n| = (n+1)|w|$.*

Proof. Condition (i) is easily proved. Indeed, for any $n > 0$, one has

$$\epsilon \Rightarrow_I^n (w'a)^n b^{kn} = (w'a)^{n-1}(w'a)b^{kn} \Rightarrow_I (w'a)^{n-1}w(w'a)b^{kn} = x_n.$$

By condition (i), for any $n > 0$, $\epsilon \Rightarrow_I^{n+1} x_n$ which yields condition (ii). □

Corollary 3.3. *Let n, m be positive integers. If $x_n \Rightarrow_I^\ell x_{n+m}$ then $\ell = m$.*

Proof. By condition (ii) of the previous lemma, $|x_{n+m}| = (n+m+1)|w| = |x_n| + m|w|$, which implies that the length of the derivation $x_n \Rightarrow_I^\ell x_{n+m}$ is $\ell = m$. □

Lemma 3.4. *Let y be a word and let n, ℓ be positive integers. If $x_n \Rightarrow_I^\ell y$ then*

i. *$y = y'b^h$ where $y' \notin A^*b$ and $1 \leq h \leq k(n+\ell)$;*
ii. *if $h = k(n+\ell)$ then $y = (w'a)^{n-1}w(w'a)^{\ell+1}b^h$.*

Proof. The claim of the lemma is easily proved by induction on the integer $\ell \geq 1$ such that $x_n \Rightarrow_I^\ell y$. □

Proposition 3.5. *Let I be a set of words satisfying the hypotheses of the first case. The derivation relation $\Rightarrow_I^*$ is not a wqo on L_I^ϵ.*

Proof. We prove that the sequence $\{x_n\}_{n>0}$ is bad. Suppose, by contradiction, that it is good. Hence, there exist positive integers n, ℓ such that $x_n \Rightarrow_I^* x_{n+\ell}$. By Corollary 3.3, the previous derivation has length ℓ, hence

$$x_n \Rightarrow_I^\ell x_{n+\ell}.$$

Since $b^{k(n+\ell)}$ is a suffix of $x_{n+\ell}$, Lemma 3.4 gives

$$x_{n+\ell} = (w'a)^{n-1}w(w'a)^{\ell+1}b^{k(n+\ell)}.$$

On the other hand

$$x_{n+\ell} = (w'a)^{n+\ell-1}w(w'a)b^{k(n+\ell)},$$

and the two factorizations above give

$$(w'a)^{\ell}w = (w'a)^{\ell}w'ab^k = w(w'a)^{\ell}.$$

This implies that $a = b$ which is a contradiction. Hence, $\Rightarrow_I^*$ is not a wqo on L_I^ϵ. □

3.2. Second case

We suppose that $\mathrm{Card}(I) \geq 2$ and, for every letter a of $\mathrm{alph}(I)$, there exists a word of I which begins with a. Set $\mathrm{alph}(I) = A$.

Lemma 3.6. *Let I be a finite avoidable set of A^+. Then there exists a word $w \in A^+$ such that, for any $n \geq 0$, $\mathrm{Fact}(w^n) \cap I = \emptyset$.*

Proof. Let $X = A^* \setminus A^*IA^*$. Since I is finite, $X \in Rec(A^*)$. Moreover, since I is avoidable in A^*, X is infinite. By the latter two conditions and by using the well known Pumping Lemma for recognizable languages, one has that there exists a word $v = fwg \in X$ with $f, g, w \in A^*$ such that $w \neq \epsilon$ and, for any $n \geq 0$, $fw^ng \in X$. Since X is closed by factors, we have that, for any $n \geq 0$, $w^n \in X$ and, thus, $\mathrm{Fact}(w^n) \cap I = \emptyset$. □

From now on, w denotes the word defined in the statement of Lemma 3.6.

Lemma 3.7. *For any $a \in A$ there exist words $ax, ay \in L_I^\epsilon$ such that $x \notin \mathrm{Suff}(y)$ and $|x| < |y|$.*

Proof. First suppose that there exists a word u of I of minimal period at least two. Hence, $u = u'cd^k$ with $u' \in A^*$, $c, d \in A, c \neq d$ and $k > 0$. Then $\epsilon \Rightarrow_I^2 (u'c)^2d^{2k}$. Let $ax = avu$ and $ay = av(u'c)^2d^{2k}$ with $av \in I$. Thus, ax and ay satisfy the claim.

If every word of I has period 1, then there exist words $a^i, b^j \in I$, $a \neq b$. Hence, take $ax = a^i$ and $ay = a^ib^j$. □

Now it is convenient to recall that, by hypothesis, for every $a \in A$, $I \cap aA^* \neq \emptyset$. Hence, there exists a word $z \in A^+$ such that $\epsilon \Rightarrow_I^* wz$. Therefore, the sequence of words $\{z^n\}$ is such that

$$\forall\, n \geq 1, \quad \epsilon \Rightarrow_I^* w^n z^n. \tag{1}$$

Let us denote $\{z_n\}$ a sequence of words of A^+ such that, for any $n > 0$, equation (1) holds if one replaces z^n with z_n and such that z_n is of minimal length.

Lemma 3.8. *The sequence $\{|z_n|\}$ is not upper bounded.*

Proof. By contradiction, suppose that our sequence is upper bounded. Thus there exists a positive integer M such that, for any $n > 0$, $|z_n| < M$. For any $n > 0$, let l_n be the length of the derivation $\epsilon \Rightarrow_I^{l_n} w^n z_n$. Since, for any $n > 0$, $\mathrm{Fact}(w^n) \cap I = \emptyset$, then $l_n < M$ and, hence, $|w^n| < MN$, where N is the maximal length of a word of I. The latter inequality is not possible if $n > MN$. Hence, the sequence $\{|z_n|\}$ is not upper bounded. □

Now let a be a letter such that $w \notin aA^*$. Consider the words ax, ay satisfying the statement of Lemma 3.7 and the sequence $\{z_n\}$ previously defined.

By possibly replacing the sequence $\{w^n z_n\}$ with one of its subsequence, Lemma 3.8 yields the following corollary.

Corollary 3.9. *The sequence of words $\{z_n\}$ is such that, for any $n, m > 0$, $|z_n| + |y| < |z_{n+m}|$.*

We consider the following two sequences $\{x_n\}$, $\{y_n\}$ of words: for any $n > 0$,

$$x_n = w^n a x z_n, \quad y_n = w^n a y z_n.$$

The condition that, for any $n > 0$, $x_n, y_n \in L_I^\epsilon$ immediately follows from the definition of the sequences $\{x_n\}$ and $\{y_n\}$. The following Lemma is used in the sequel. Its proof is an easy consequence of the definition of the relation $\Rightarrow_I^*$.

Lemma 3.10. *Let $f, g, v \in A^*$ and let $a \in A$. If $fag \Rightarrow_I^* v$ then $v = f'ag'$ where f', g' are words of A^* such that: $f \Rightarrow_I^* f'$ and $g \Rightarrow_I^* g'$.*

Lemma 3.11. *Let n, k be positive integers. If $x_n \Rightarrow_I^* x_{n+k}$ then $z_{n+k} = z' x z_n$, $z' \in A^*$. Similarly, if $y_n \Rightarrow_I^* y_{n+k}$ then $z_{n+k} = z'' y z_n$, $z'' \in A^*$.*

Proof. We deal with the first case, that is, $x_n \Rightarrow_I^* x_{n+k}$, the other case being completely analogous. By applying Lemma 3.10 to $f = w^n$ and $g = x z_n$ one obtains words $f', g' \in A^*$ such that

1. $f'ag' = x_{n+k}$;
2. $w^n \Rightarrow_I^* f'$;
3. $x z_n \Rightarrow_I^* g'$.

First we remark that if $f' = w^n$, then by (1) $w \in aA^*$, which is not possible since w does not begin with the letter a. Hence, by (2), $w^n \Rightarrow_I^+ f'$. This implies that there exists at least one word $u \in I$ such that $u \in \mathrm{Fact}(f')$. If $|f'| \leq |w^{n+k}|$ then, by condition (1), $f' \in \mathrm{Pref}(w^{n+k})$ so that $u \in \mathrm{Fact}(w^{n+k})$. By Lemma 3.6 the latter condition is not possible. Hence, $|f'| > |w^{n+k}|$ so that, by condition (1), $f' = w^{n+k}\zeta$, where $\zeta \in A^+$. Since $\epsilon \Rightarrow_I^* w^n z_n$, condition (2) yields $\epsilon \Rightarrow_I^* w^n z_n \Rightarrow_I^* f' z_n = w^{n+k}\zeta z_n$. Hence, by the definition of z_{n+k}, $|\zeta z_n| \geq |z_{n+k}|$, so that $|w^{n+k}\zeta a x z_n| \geq |w^{n+k} a x z_{n+k}|$. On the other hand, we have

$$x_n = w^n a x z_n \Rightarrow_I^* w^{n+k}\zeta a x z_n \Rightarrow_I^* w^{n+k} a x z_{n+k} = x_{n+k},$$

which gives $w^{n+k}\zeta a x z_n = w^{n+k} a x z_{n+k}$. Hence, by Corollary 3.9, $z_{n+k} = z' x z_n$, with $z' \in A^*$. □

Proposition 3.12. *Let I be a set of words satisfying the hypotheses of the second case. The relation $\Rightarrow_I^*$ is not a wqo on L_I^ϵ.*

Proof. The proof is by contradiction. Set $L = L_I^\epsilon$ and denote $\leq$ the relation $\Rightarrow_I^*$. Suppose that L is well quasi-ordered by $\leq$. Then, by Lemma 2.2, the set $L \times L$ is well quasi-ordered by the canonical relation defined by $\leq$ on $L \times L$. Hence, every sequence of elements of $L \times L$ is good with respect to that quasi-order. Now consider the sequence $\{(x_n, y_n)\}$. Hence there exist integers $n, k > 0$ such that $x_n \leq x_{n+k}$ and $y_n \leq y_{n+k}$. By Lemma 3.11, $z_{n+k} = z'xz_n = z''yz_n$ with z', $z'' \in A^*$. On the other hand, by Lemma 3.7, $|x| < |y|$ and therefore, x is a suffix of y. Again, by Lemma 3.7, this is a contradiction. Hence, $\Rightarrow_I^*$ is not a wqo on L_I^ϵ. □

Remark 3.13. The same result of Proposition 3.12 may be obtained under the assumption that, for every letter $a \in A$, there exists a word of the set I that ends with a. In this case, the proof is completely analogous.

3.3. Third case

Now we suppose that the set I has at least two words and it does not satisfy the hypothesis of Case 2. Set $\mathrm{alph}(I) = A$. Therefore, according to Remark 3.13, we assume that there exists a letter c of the set A such that, for every $f \in I$, $f \notin A^*c$. In this case we also suppose that at least one word of I begins with a letter $a \neq c$. In order to study this case, it is useful to introduce some preliminary definitions and results. For any $f \in A^+$, we set

$$\nu_c(f) = \frac{|f|_c}{|f|}.$$

If f is the empty word we set $\nu_c(f) = 0$. We adopt the following conventions. The word u denotes a prefix of a word of the set I such that $\nu_c(u)$ is maximal. Moreover, w denotes a word of the set I with $u \in \mathrm{Pref}(w)$ and we set $w = uv, v \in A^*$.

Lemma 3.14. *The following conditions hold.*

i. *The word u ends with the letter c and $v \neq \epsilon$.*
ii. *Let f be a word of I such that, for any $h \in I$, $\nu_c(h) \leq \nu_c(f)$. Then, for any $g \in L_I^\epsilon$, $\nu_c(g) \leq \nu_c(f)$. Moreover, $\nu_c(f) < \nu_c(u)$.*
iii. *Let $n > 0$ and let f be a word of A^* such that $u^n \Rightarrow_I^* f$. Then $\nu_c(f) \leq \nu_c(u)$.*
iv. *Let $v_0, \ldots, v_i$ be words of the set $\mathrm{Pref}(L_I^\epsilon)$. Then $\nu_c(v_0 \cdots v_i) \leq \nu_c(u)$.*

Proof. *i*) If u does not end with c then $u = u'a$, with $a \neq c$ and thus $\nu_c(u) < \nu_c(u')$ which contradicts the choice of u. Since $w \notin A^*c$ and $u \in A^*c$, one has $v \neq \epsilon$.
ii) The first part of the claim may be easily proved by induction on the length of the derivation which yields g starting from the empty word. For the second part of (*ii*), by hypothesis, $f \notin A^*c$ and thus $f = f'a$ with $f' \in A^*, a \neq c$, whence $\nu_c(f) < \nu_c(f')$. Since f' is a prefix of a word of I, by the choice of u, we have $\nu_c(f') \leq \nu_c(u)$ and thus $\nu_c(f) < \nu_c(u)$.

iii) The claim may be easily proved by induction on the length of the derivation that yields f starting from u^n.
iv) It is enough to prove that, for any $f \in \mathrm{Pref}(L_I^\epsilon)$, $\nu_c(f) \leq \nu_c(u)$. Under this assumption, there exists a word $t \in L_I^\epsilon$ such that $f \in \mathrm{Pref}(t)$. Suppose that $\epsilon \Rightarrow_I^n t$, with $n \geq 0$. We prove the claim by induction on n. The claim is trivial if $n = 0$. If $n = 1$ then the claim follows from the choice of u. Hence, the basis of the induction is proved. Let us prove the inductive step. Then we have

$$\epsilon \Rightarrow_I^{n-1} t' \Rightarrow_I t$$

with $n > 1$. We have to examine the following cases:

1. $t' = ft_1t_2$, $t = ft_1gt_2$, with $g \in I$. Hence $f \in \mathrm{Pref}(t')$ and the claim follows by the induction step;
2. $t' = f't_1, t = f'ht_1$, where $f = f'g, h = gg' \in I$ and $g, t_1 \in A^+$. Since f' is a prefix of t', by induction hypothesis, $\nu_c(f') \leq \nu_c(u)$ and since g is a prefix of a word of I, $\nu_c(g) \leq \nu_c(u)$. Therefore, we have $\nu_c(f) = \nu_c(f'g) \leq \nu_c(u)$;
3. $t = f_1gf_2t_1$, $t' = f_1f_2t_1$, $f = f_1gf_2$, with $f_1, f_2, t_1 \in A^*$ and $g \in I$. Since f_1f_2 is a prefix of t', by induction hypothesis, $\nu_c(f_1f_2) \leq \nu_c(u)$. Since $g \in I$, by (ii), one has that $\nu_c(g) < \nu_c(u)$. Hence, the latter two conditions give $\nu_c(f) = \nu_c(f_1gf_2) < \nu_c(u)$.

□

Now it is convenient to notice that, by hypothesis, $\epsilon \Rightarrow_I w = uv$ and therefore,

$$\epsilon \Rightarrow_I^* u^n v^n. \tag{2}$$

Let us denote $\{z_n\}$ a sequence of words of A^+ such that, for any $n > 0$, equation (2) holds if one replaces v^n with z_n and such that z_n is of minimal length.

Lemma 3.15. *The sequence $\{|z_n|\}$ is not upper bounded.*

Proof. The proof is by contradiction. Suppose that our sequence is upper bounded and thus there exists a positive integer M such that, for any $n > 0$, $|z_n| < M$. Hence, we have

$$\lim_{n\to\infty} \nu_c(u^n z_n) = \nu_c(u). \tag{3}$$

Let f be a word of I such that $\nu_c(f)$ is maximal in I. Set $\delta = \nu_c(u) - \nu_c(f)$. By Lemma 3.14 – (ii), $\delta > 0$ and, for any $g \in L_I^\epsilon$,

$$\delta \leq \nu_c(u) - \nu_c(g). \tag{4}$$

On the other hand, for any $n > 0$, $u^n z_n \in L_I^\epsilon$ and, for any sufficiently large n, Equation (3) gives

$$|\nu_c(u) - \nu_c(u^n z_n)| = \nu_c(u) - \nu_c(u^n z_n) < \frac{\delta}{2},$$

which contradicts equation (4). Hence, the sequence $\{|z_n|\}$ is not upper bounded.
□

The following result is useful. Its proof is similar to that of Lemma 3.7.

Lemma 3.16. *Let $a \in A$, with $a \neq c$ and $I \cap aA^* = \emptyset$, and let $H = |w| + 1$. Then there exist words $a^H x, a^H y \in L_I^\epsilon$ such that $x \notin \mathrm{Suff}(y)$ and $|x| < |y|$.*

By possibly replacing the sequence $\{u^n z_n\}$ with one of its subsequence, Lemma 3.15 yields the following corollary.

Corollary 3.17. *The sequence of words $\{z_n\}$ is such that, for any $n, m > 0$, $|z_n| + |y| < |z_{n+m}|$ where y is the word defined in Lemma 3.16.*

Now, starting from the words $a^H x, a^H y, w = uv$ and those of the sequence $\{z_n\}$, we consider the following two sequences $\{x_n\}$, $\{y_n\}$ of words: for any $n > 0$,

$$x_n = u^n a^H x z_n, \quad y_n = u^n a^H y z_n.$$

The condition that, for any $n > 0$, $x_n, y_n \in L_I^\epsilon$ immediately follows from the definition of the sequences $\{x_n\}$ and $\{y_n\}$.

Lemma 3.18. *Let n, k be positive integers. If $x_n \Rightarrow_I^* x_{n+k}$ then $z_{n+k} = z' x z_n$, $z' \in A^+$. Similarly, if $y_n \Rightarrow_I^* y_{n+k}$ then $z_{n+k} = z'' y z_n$, $z'' \in A^+$.*

Proof. We deal with the first case, that is, $x_n \Rightarrow_I^* x_{n+k}$, the other being completely analogous. By applying Lemma 3.10 to $f = u^n$ and $g = a^{H-1} x z_n$ one obtains words $f', g' \in A^*$ such that

1. $f' a g' = u^{n+k} a^H x z_{n+k}$;
2. $u^n \Rightarrow_I^* f'$;
3. $a^{H-1} x z_n \Rightarrow_I^* g'$.

Let us prove that $u^{n+k} \in \mathrm{Pref}(f')$. By (1), it suffices to show that $|f'| \geq |u^{n+k}|$. By contradiction, suppose that $|f'| < |u^{n+k}|$. First we notice that in the derivation process

$$g = a^{H-1} x z_n \Rightarrow_I^* g'$$

at least one word of I must be inserted in the prefix a^{H-1} of g. Indeed, otherwise, we have $g' = a^{H-1} g''$, $g'' \in A^*$ and therefore

$$f' a g' = f' a^H g'' = u^{n+k} a^H x z_{n+k}.$$

Since $|f'| < |u^{n+k}|$ and $|u| < H$ we obtain $u \in A^* a$, with $a \neq c$ which contradicts condition (i) of Lemma 3.14. Therefore, the prefix of ag' of length $H - 1$ is of the form

$$p = a v_1 \cdots a v_i,$$

where $1 \leq i \leq H - 1$, $v_1, \ldots, v_i \in \mathrm{Pref}(L_I^\epsilon)$. Again, the equality $f' a g' = u^{n+k} a^H x z_{n+k}$ and the condition $|f'| < |u^{n+k}|$, $|u| < H$ yield the existence of a power u^j of u such that $j \leq n + k$ and

$$u^j = f' q,$$

where q is a proper prefix of p. Set $q = av_1 \cdots av'_k, v'_k \in \mathrm{Pref}(v_k)$. Then, by Lemma 3.14 – (iv), we have $\nu_c(q) = \nu_c(av_1 \cdots av'_k) < \nu_c(v_1 \cdots v'_k) \leq \nu_c(u)$ and by Lemma 3.14 – (iii) $\nu_c(f') \leq \nu_c(u)$ whence

$$\nu_c(u) = \nu_c(u^j) = \nu_c(f'q) < \nu_c(u),$$

which is a contradiction. Hence, $|f'| \geq |u^{n+k}|$ and thus by (1), $f' = u^{n+k}\zeta$, $\zeta \in A^*$. Therefore, we have $f = u^n \Rightarrow_I^+ f' = u^{n+k}\zeta$. On the other hand, we have $\epsilon \Rightarrow_I^* u^n z_n$ which thus gives

$$\epsilon \Rightarrow_I^* u^{n+k}\zeta z_n.$$

By the definition of z_{n+k}, we have $|\zeta z_n| \geq |z_{n+k}|$ and thus

$$|u^{n+k}\zeta a^H x z_n| \geq |u^{n+k} a^H x z_{n+k}|.$$

Now, by Lemma 3.10,

$$fag = u^n a^H x z_n \Rightarrow_I^+ f'ag = u^{n+k}\zeta a^H x z_n \Rightarrow_I^* u^{n+k} a^H x z_{n+k} = f'ag',$$

which gives $u^{n+k}\zeta a^H x z_n = u^{n+k} a^H x z_{n+k}$. By Corollary 3.17, $|xz_n| < |z_{n+k}|$ which gives $z_{n+k} = z'xz_n$, with $z' \in A^+$. □

The proof of the following proposition follows *verbatim* the argument of that of Proposition 3.12.

Proposition 3.19. *Let I be a set of words satisfying the hypotheses of the third case. The relation $\Rightarrow_I^*$ is not a wqo on L_I^ϵ.*

3.4. Fourth case

Finally we suppose that the set I has at least two words, its alphabet is $A = \{a, b\}$ and $I \subseteq aA^*b$. One can easily show that this case must necessarily occur if all the previous cases (and the symmetric ones) are excluded. In order to study this case, we need some preliminary results. For any $f \in aA^*b$, we can write $f = a^k f'$, with $k \geq 1$ and $f' \in bA^*$. We set

$$\mu(f) = k.$$

From now on, the word $w = a^k v, v \in bA^*$ denotes a word of I such that the ratio $\mu(w)/|w|$ is maximal. Since $w = a^k v \in I$, for any $n > 0$, one has

$$\epsilon \Rightarrow_I^n a^{kn} v^n = a^{k(n-1)} a^{k-1} \cdot av^n \Rightarrow_I a^{k(n-1)} a^{k-1} \cdot a^k v \cdot av^n = a^{kn} a^{k-1} vav^n. \quad (5)$$

For every $n > 0$, we set

$$x_n = a^{kn} a^{k-1} vav^n.$$

By (5), all words x_n belong to the language L_I^ϵ.

Lemma 3.20. *Let $f = a^p f'$ with $p > 0$, $f' \in bA^*$ and let g be a word such that $f \Rightarrow_I^* g$. If*

$$\mu(g) = p + k\ell \ \text{ and } \ |g| = |f| + |w|\ell,$$

then

$$g = a^{p+k\ell} x f',$$

with $x \in A^*$.

Proof. By hypothesis we have that

$$f \Rightarrow_I^m g, \text{ with } m \geq 1. \tag{6}$$

Let $u_1, \ldots, u_m$ be the m words of I used in the derivation process above and, for every $i = 1, \ldots m$, set $p_i = \mu(u_i)$ and $q_i = |u_i|$. By hypothesis

$$\sum_{i=1,\ldots,m} q_i \ = \ |w|\ell. \tag{7}$$

By the choice of w, since $\mu(w)/|w| = k/|w|$ is maximal, for every $i = 1, \ldots, m$, we have

$$p_i \leq \frac{q_i k}{|w|},$$

and thus

$$\sum_{i=1,\ldots,m} p_i \ \leq \ \frac{k}{|w|} \sum_{i=1,\ldots,m} q_i.$$

Hence, by (7),

$$\sum_{i=1,\ldots,m} p_i \ \leq \ \ell k. \tag{8}$$

Let $\alpha \Rightarrow_I \beta$ be the i-th production of the derivation process (6), so that β is obtained from α by the insertion in α of the word u_i. The insertion is called *useful* if it is done immediately after the prefix $a^{\mu(\alpha)}$ of α, that is $\beta \in a^{\mu(\alpha)+p_i} bA^*$. It is clear that if the insertion is not useful, then $\mu(\beta) < \mu(\alpha) + p_i$.

Now suppose that there exists at least one production in (6) such that the corresponding insertion is not useful. By the previous argument and by (8), we have

$$\mu(g) < \mu(f) + \sum_{i=1,\ldots,m} p_i \ \leq \ \mu(f) + k\ell,$$

and this contradicts the assumption on g. Therefore all the insertions in (6) are useful and this implies that the prefix f' is preserved in the whole derivation process. □

Corollary 3.21. *Let n, ℓ be two positive integers. If $x_n \Rightarrow_I^* x_{n+\ell}$ then*

$$x_{n+\ell} = a^{k(n+\ell)} a^{k-1} x a v^n,$$

with $x \in A^*$.

Proof. By the hypothesis and the fact that

$$\mu(x_{n+\ell}) = \mu(x_n) + k\ell, \quad |x_{n+\ell}| = |x_n| + \ell|w|,$$

the claim follows by applying Lemma 3.20 to $f = x_n$ and $g = x_{n+\ell}$. □

Proposition 3.22. *Let I be a set of words satisfying the hypothesis of the fourth case. Then the relation $\Rightarrow_I^*$ is not a wqo on L_I^ϵ.*

Proof. We prove that the sequence $\{x_n\}_{n>0}$ is bad. Suppose, by contradiction, that it is good. Then there exist integers $n, \ell > 0$ such that $x_n \Rightarrow_I^* x_{n+\ell}$. By Corollary 3.21,

$$x_{n+\ell} = a^{k(n+\ell)}a^{k-1}xav^n,$$

whereas, by definition of $x_{n+\ell}$, one has

$$x_{n+\ell} = a^{k(n+\ell)}a^{k-1}vav^{n+\ell}.$$

By the latter two factorizations of $x_{n+\ell}$ one has that the word v ends with the letter $a \neq b$ which is a contradiction. Therefore, $\{x_n\}$ is a bad sequence in L_I^ϵ and $\Rightarrow_I^*$ is not a well quasi order on L_I^ϵ. □

By Theorems 2.3 and 3.1 and Lemma 2.2, we have that $\Rightarrow_I^*$ is a well quasi-order on A^* if and only if $\Rightarrow_I^*$ is a well quasi-order on L_I^ϵ. Hence, we obtain the following corollary.

Corollary 3.23. *Let I be a finite set of words. Then the following conditions are equivalent:*

i. *$\Rightarrow_I^*$ is a well quasi-order on L_I^ϵ;*
ii. *I is unavoidable;*
iii. *L_I^ϵ is regular.*

4. Well quasi-orders and shuffle

As announced in the introduction of this paper, one can consider a possible extension of the previous results with respect to other significant quasi-orders and, in particular, in the case of the relation $\vdash_I^*$ whose definition is recalled below. 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 $L_{\vdash_I}^\epsilon$ denotes the set of all words derived from the empty word by applying $\vdash_I^*$, that is,

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

The relation $\vdash_I^*$ has been considered in [9] where the following theorem has been proved.

Theorem 4.1. *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 $|u| \geq k$, contains as a subsequence a word of I;*
iii. *the language $L^\epsilon_{\vdash_I}$ is regular.*

In [9] it is also proved that I is subsequence unavoidable if and only if, for every $a \in A$, $I \cap \{a\}^+ \neq \emptyset$. It is also worth noticing that the relationships between the quasi-orders $\vdash_I^*$ and $\Rightarrow_I^*$ have been deeply investigated in [2], [3] where, as a consequence of a more general result, the following theorem is proved:

Theorem 4.2. *For any finite set I, $\vdash_I^*$ is a wqo on L_I^ϵ.*

In this theoretical setting, it is natural to ask whether Theorem 4.1 may be extended by replacing condition (i) with the weaker condition that the derivation relation $\vdash_I^*$ is a wqo on $L^\epsilon_{\vdash_I}$. Unfortunately this is not true as shown by the following example. Consider the set $I = \{ab\}$. It is easily verified that $L^\epsilon_{\vdash_I} = L_I^\epsilon$ and therefore, by a well-known construction, $L^\epsilon_{\vdash_I}$ is generated by a context-free grammar with only one variable. Precisely, $L^\epsilon_{\vdash_I}$ is the language of all *semi-Dyck words* over the alphabet $\{a, b\}$. By Theorem 4.2, $\vdash_I^*$ is a well quasi-order on $L^\epsilon_{\vdash_I} = L_I^\epsilon$ while this language is not regular. This example leads us to further investigate the relationships between the wqo property of $\vdash_I^*$ and the context-freeness of the language $L^\epsilon_{\vdash_I}$.

We conjecture that $\vdash_I^*$ is a wqo on $L^\epsilon_{\vdash_I}$ if $L^\epsilon_{\vdash_I}$ is context-free. It seems that a significant step of a possible solution of our problem is the combinatorial characterization of finite sets I such that $L^\epsilon_{\vdash_I}$ is context-free. In the literature, the language $L^\epsilon_{\vdash_I}$ is also called the *iterated shuffle of I* or *the shuffle closure of I* [13]. Many papers have been devoted to the studying of the shuffle closure of finite languages (see for instance [13, 14]) but, as far as we know, no characterization has been given for the context-freeness property of them. Now we give such a characterization when I is a singleton. In order to prove this result, we need the following lemma.

Lemma 4.3. *Let $I = \{w\}$ with $w = a_1^{i_1} a_2^{i_2} \cdots a_k^{i_k}$ where $k \geq 3$, $i_1, \ldots, i_k \geq 1$ and the $a_i's$ are letters such that, for every $i = 1, \ldots, k-1$, $a_i \neq a_{i+1}$.*
If $X = \{a_1^{i_1 n} a_2^{i_2 n} \cdots a_k^{i_k n} \mid n \geq 1\}$ and $R = (a_1^{i_1})^*(a_2^{i_2})^* \cdots (a_k^{i_k})^*$, then

$$X = L^\epsilon_{\vdash_I} \cap R.$$

Proof. The inclusion $X \subseteq L^\epsilon_{\vdash_I} \cap R$ is easily proved. Let $x = a_1^{ni_1} a_2^{ni_2} \cdots a_k^{ni_k}$, with $n \geq 1$. Obviously, $x \in R$. On the other hand, one easily verifies that $\epsilon \vdash_I^n x$ and hence, $x \in L^\epsilon_{\vdash_I}$. This proves that $x \in L^\epsilon_{\vdash_I} \cap R$.

Let us now prove that $X \supseteq L^{\epsilon}_{\vdash_I} \cap R$. Let $x \in L^{\epsilon}_{\vdash_I}$, $x \neq \epsilon$, and let $n(x)$ be the number of distinct powers of letters of alph(w) which factorize x. It is useful to remark that $n(w) = k$.

Let $\epsilon \vdash_I x_1 = w \vdash_I x_2 \vdash_I \cdots \vdash_I x_n = x$ be the sequence of derivations which yields x. The following properties may be easily proved by induction on n:
– for every $\ell = 1, \ldots, n$, $n(x_\ell) \geq k$ and, if $n(x_\ell) > k$, then $x \notin R$;
– the sequence of integers $\{n(x_\ell)\}_{\ell=1,\ldots n}$ is monotone non decreasing.

By using the two properties above, one may easily prove that, if $x \in L^{\epsilon}_{\vdash_I} \cap R$, then $n(x) = k$ and $x = a_1^{ni_1} a_2^{ni_2} \cdots a_k^{ni_k}$, where n is the length of the sequence of derivations of x. This proves that $x \in X$. □

Theorem 4.4. *Let $I = \{w\}$. Then $L^{\epsilon}_{\vdash_I}$ is context-free if and only if $w = a^k b^h$ where a and b are distinct letters and h, k are non negative integers.*

Proof. Let us prove the necessary condition. Let

$$I = \{w\} = \{a_1^{i_1} a_2^{i_2} \cdots a_k^{i_k}\},$$

where $k \geq 3$, $i_1, \ldots, i_k \geq 1$ and the $a_i's$ are letters such that, for every $i = 1, \ldots, k-1$, $a_i \neq a_{i+1}$. By contradiction, suppose that $L^{\epsilon}_{\vdash_I}$ is context-free. Let

$$X = \{a_1^{i_1 n} a_2^{i_2 n} \cdots a_k^{i_k n} \mid n \geq 1\}.$$

It is well-known that X is not context-free and this result may be proved by applying the Pumping Lemma for context-free languages to X. On the other hand, Lemma 4.3 gives

$$X = L^{\epsilon}_{\vdash_I} \cap (a_1^{i_1})^* (a_2^{i_2})^* \cdots (a_k^{i_k})^*.$$

Since the family of context-free languages is closed under intersection with regular languages, one has that X is context-free which is a contradiction. This proves the necessary condition.

Let us now prove the sufficient condition. Let $I = \{a^h b^k\}$, where a and b are distinct letters and h, k are non negative integers. If $k = 0$ (resp. $h = 0$), then $L^{\epsilon}_{\vdash_I} = (a^h)^*$ (resp. $= (b^k)^*$) and, hence, it is regular. Suppose that $h, k > 0$. For any word u over the alphabet $\{a, b\}$, one can consider the following integer parameters

$$q_a^u = |u|_a / h, \quad q_b^u = |u|_b / k, \quad \text{and}$$
$$r_a^u = |u|_a \bmod h, \quad r_b^u = |u|_b \bmod k.$$

Now we prove that for any word w,

$$w \in L^{\epsilon}_{\vdash_I}$$

if and only if the following condition holds:

(A) $q_a^w = q_b^w$, $r_a^w = r_b^w = 0$ and for any prefix u of w, either $q_a^u > q_b^u$ or $q_a^u = q_b^u$ and $r_b^u = 0$.

In order to prove the above characterization let us denote by L the language of all the words satisfying condition **(A)**. We have to prove that $L^{\epsilon}_{\vdash_I} = L$. Obviously the empty word belongs to L. Moreover, one can easily check that if $u \in L$ and $u \vdash_I v$, then $v \in L$. This shows that $L^{\epsilon}_{\vdash_I} \subseteq L$.

To gain the inverse inclusion we prove that for any non empty word $w \in L$ there exists a word $w' \in L$ such that $w' \vdash_I w$. By proceeding on induction on the length of any word w, this fact gives $w \in L$ whenever $w \in L^{\epsilon}_{\vdash_I}$. Then let w be a non empty word in L. Since w satisfies condition **(A)**, w must contain as a prefix the word a^h and at least k occurrences of the letter b. Consider the first k occurrences of b and write w as

$$w = a^h a^{j_1} b a^{j_2} b \dots a^{j_k} b v.$$

Let $w' = a^{j_1} a^{j_2} \dots a^{j_k} v$. Obviously $w' \vdash_I w$. Moreover, since $|w'|_a = |w|_a - h$ and $|w'|_b = |w|_b - k$, one has $q_a^{w'} = q_a^w - 1$, $q_b^{w'} = q_b^w - 1$, $r_a^{w'} = r_a^w$, and $r_b^{w'} = r_b^w$. Therefore, $q_a^{w'} = q_b^{w'}$, and $r_a^{w'} = r_b^{w'} = 0$, *i.e.*, w' satisfies the first part of condition **(A)**. Now let u' be a prefix of w'. If u' is a prefix of $a^{j_1} a^{j_2} \dots a^{j_k}$, then it trivially satisfies the second part of condition **(A)**. If u' is not a prefix of $a^{j_1} a^{j_2} \dots a^{j_k}$, then it can be written as $u' = a^{j_1} a^{j_2} \dots a^{j_k} \lambda$, where λ is a prefix of v. Then, consider the word $u = a^h a^{j_1} b a^{j_2} b \dots a^{j_k} b \lambda$. Observe that u is a prefix of w and, by hypothesis it satisfies the second part of **(A)**. Since $|u'|_a = |u|_a - h$ and $|u'|_b = |u|_a - k$, one has $q_a^{u'} = q_a^u - 1$, $q_b^{u'} = q_b^u - 1$, $r_a^{u'} = r_a^u$, and $r_b^{u'} = r_b^u$. The above equalities imply that u' must satisfy the second part of condition **(A)**.

By using the characterization above, one may construct a push-down automaton (PDA) M that accepts $L^{\epsilon}_{\vdash_I}$. We give a short description of M. The PDA M uses the stack as a counter. The stack alphabet contains only two letters A and Z_0. The letter Z_0 is the *start symbol* of M and, in any computation, it will appear only in the bottom of the stack. The letter A is used to give a unary representation of an integer, as explained below. During the computation, any time the number of a increases of h then the counter increases of 1 by adding a letter A to the top of the stack. When the number of b increases of k, then the counter decreases of 1 by erasing the letter A from the top of the stack. The finite states are used to count the a's modulo h and the b's modulo k. In such a way the PDA M, after reading a prefix u, can compute in its stack the (non negative) integer $q_a^u - q_b^u$, and, in its finite states, the integers r_a^u and r_b^u. We remark that M may decide whether $q_a^u > q_b^u$ or $q_a^u = q_b^u$, by checking whether the symbol on top of the stack is A or Z_0. The details on the construction of M and how it accepts the words satisfying condition **(A)**, are left to the reader. □

The most simple case where $I = \{w\}$ is not context-free happens when $w = abc$. In this case we are able to prove that $\vdash_I^*$ is not a wqo on $L^{\epsilon}_{\vdash_I}$. In order to achieve this result, we need to recall some definitions and results stated in [2].

Definition 4.5. Let $u = a_1 \cdots a_n$ and $v = b_1 \cdots b_m$ be two words over A with $n \leq m$. An *embedding* of u in v is a map $f : \{1, \dots, n\} \longrightarrow \{1, \dots, m\}$ such that f is increasing and, for every $i = 1, \dots, n$, $a_i = b_{f(i)}$.

It is useful to remark that a word u is a subsequence of v if and only if there exists an embedding of u in v.

Definition 4.6. Let $u, v \in A^*$ 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

$$\langle v - u \rangle_f = b_{i_1} \cdots b_{i_\ell}$$

where $\{i_1, i_2, \ldots, i_\ell\}$ is the increasing sequence of all the integers of $\{1, \ldots, m\}$ not belonging to $\mathrm{Im}(f)$. 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.

Moreover, 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 $a_i \in A$, $v_i \in A^*$.

Lemma 4.7. *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. The proof is by induction. By hypothesis there exists $k \geq 0$ 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 $u = u_1 u_2 \cdots u_{m+1}$ and $v = u_1 b_1 u_2 b_2 \cdots b_m u_{m+1}$ with $b_1 b_2 \cdots b_m \in I$. The previous factorizations of u and v yield an embedding f of u in v such that $\langle v - u \rangle_f = b_1 b_2 \cdots b_m \in L^{\epsilon}_{\vdash_I}$. 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 that the embedding f generates the factorizations $u = a_1 \cdots a_n$ and $w = u_1 a_1 u_2 a_2 \cdots u_i a_i \cdots u_n a_n u_{n+1}$ with $a_i \in A$, $u_i \in A^*$. Hence, $\langle w - u \rangle_f = u_1 u_2 \cdots u_{n+1} \in L^{\epsilon}_{\vdash_I}$.

Since $w \vdash_I v$, there exists a word $b_1 b_2 \cdots b_m \in I$ such that v is obtained from w inserting the sequence of letters $b_1, b_2 \cdots, b_m$. Since the word w is factorized as $w = u_1 a_1 u_2 a_2 \cdots u_i a_i \cdots u_n a_n u_{n+1}$, the word v can be factorized as $v = u'_1 a_1 u'_2 a_2 \cdots u'_i a_i \cdots u'_n a_n u'_{n+1}$ where each u'_i is obtained from u_i inserting some of the letters of $b_1 b_2 \cdots b_m$. The above factorization of v gives an embedding g of u in v such that

$$\langle v - u \rangle_g = u'_1 u'_2 \cdots u'_{n+1}.$$

Moreover, as $u_1 u_2 \cdots u_{n+1} \vdash_I u'_1 u'_2 \cdots u'_{n+1}$, one has $\langle v - u \rangle_g \in L^{\epsilon}_{\vdash_I}$. □

The following lemma is useful and its proof is left to the reader.

Lemma 4.8. *Let $I = \{w\}$ with $w = abc$ and let $u \in \{a, b, c\}^*$. Then $u \in L^{\epsilon}_{\vdash_I}$ if and only if $|u|_a = |u|_b = |u|_c$ and, for every $p \in \mathrm{Pref}(u)$, $|p|_a \geq |p|_b \geq |p|_c$.*

Let $A = \{a, b, c\}$ and consider the sequence $\{S_n\}_{n\geq 1}$ of words of A^* defined as: for every $n \geq 1$,

$$S_n = a(a^2b^2c^2)(bac)^n b(a^2b^2c^2)c.$$

The following result holds.

Proposition 4.9. *The sequence $\{S_n\}_{n\geq 1}$ is bad with respect to $\vdash_I^*$ and its elements are in $L^\epsilon_{\vdash_I}$. Therefore, $\vdash_I^*$ is not a wqo on $L^\epsilon_{\vdash_I}$.*

Proof. One can easily check that, for any $n \geq 1$, $S_n \in L^\epsilon_{\vdash_I}$. Let us prove that the sequence $\{S_n\}$ is bad. By contradiction, assume that $\{S_n\}$ is good so that there exist positive integers n, ℓ such that $S_n \vdash_I^* S_{n+\ell}$. Thus, by Lemma 4.7, there exists an embedding f of S_n into $S_{n+\ell}$ such that $\langle S_{n+\ell} - S_n\rangle_f \in L^\epsilon_{\vdash_I}$.

Remark that

$$S_n = P(a^2b^2c^2)c, \qquad S_{n+\ell} = P(acb)^\ell(a^2b^2c^2)c$$

with $P = a(a^2b^2c^2)(bac)^n b$. Now we prove the following two steps.

Step 1. The embedding f is the identity on the prefix $a(a^2b^2c^2)$ of P.
First remark that $f(2) \geq 2$. Let us prove that $f(2) = 2$. Indeed, otherwise, supposing $f(2) = k > 2$ implies $f(3) \geq k + 3$ and $f(1) < k$. Then one can easily check that $\langle S_{n+\ell} - S_n\rangle_f$ admits a prefix q such that $|q|_b > |q|_a$. Hence, by Lemma 4.8, $\langle S_{n+\ell} - S_n\rangle_f \notin L^\epsilon_{\vdash_I}$ and this contradicts the choice of f. Therefore, $f(2) = 2$ and $f(1) = 1$. Again, by Lemma 4.8, the previous two equalities and the fact that $\langle S_{n+\ell} - S_n\rangle_f \in L^\epsilon_{\vdash_I}$ imply that f is the identity on $a(a^2b^2c^2)$.

Step 2. The embedding f is the identity on P.
By contradiction, suppose that the statement of the step is false and let $p\sigma$, $\sigma \in A$, be the shortest prefix of P such that f is not the identity on $p\sigma$. By Step 1, $|a(a^2b^2c^2)| \leq |p| < |P|$. Let $i = |p| - |a(a^2b^2c^2)|$ mod. 3. If $i = 0$ (resp. $i = 1$, $i = 2$), then $\sigma = b$ (resp. $\sigma = a$, $\sigma = c$). Since f is the identity on p and $f(|p| + 1) > |p| + 1$, $\langle S_{n+\ell} - S_n\rangle_f$ is a word of the set bA^* (resp. acA^*, cbA^*). By Lemma 4.8, $\langle S_{n+\ell} - S_n\rangle_f \notin L^\epsilon_{\vdash_I}$ and this contradicts the choice of f. Hence, f is the identity on P.

Since $Paa \in \mathrm{Pref}(S_n)$, Step 2 gives $f(|P|+2) > |P|+2$ and, thus, $\langle S_{n+\ell} - S_n\rangle_f \in \{ac,\ c\}A^+$. Hence, by Lemma 4.8, $\langle S_{n+\ell} - S_n\rangle_f \notin L^\epsilon_{\vdash_I}$ and this contradicts the choice of f. Therefore, the condition $S_n \vdash_I^* S_{n+\ell}$ does not hold and the sequence $\{S_n\}_{n\geq 1}$ is bad. □

Let us now consider the case when $I = \{a^h b\}$ (or symmetrically $I = \{ab^h\}$), with $h \geq 1$. Note that, by Theorem 4.4, the set $L^\epsilon_{\vdash_I}$ is context-free. The following proposition holds:

Proposition 4.10. *Let $I = \{a^h b\}$, then $L^\epsilon_{\vdash_I} = L^\epsilon_I$.*

Proof. We have to prove that $L^\epsilon_{\vdash_I} \subseteq L^\epsilon_I$, since the inverse inclusion trivially follows from the definition of $L^\epsilon_{\vdash_I}$ and L^ϵ_I.

For any word u define

$$q_a^u = |u|_a/h, \;\; r_a^u = |u|_a \bmod h.$$

Using the characterization of the set $L^\epsilon_{\vdash_I}$ given in the proof of Theorem 4.4, we can state that a word w is in $L^\epsilon_{\vdash_I}$ if and only if

(A) $q_a^w = |w|_b$, $r_a^w = 0$ and for any prefix u of w one has $q_a^u \geq |u|_b$.

Let w be a word of $L^\epsilon_{\vdash_I}$ and proceed by induction on its length. If w is empty, then $w \in L^\epsilon_I$. Suppose now that w is non empty. Since, by construction, w starts with a block a^h, then it can be written as $w = a^k a^h b\lambda$, with $k \geq 0$ and $\lambda \in \{a, b\}^*$. Let $v = a^k\lambda$. Since w satisfies the above condition **(A)**, an easy computation shows that also the word v satisfies **(A)** and, thus, $v \in L^\epsilon_{\vdash_I}$. By induction $v \in L^\epsilon_I$. Moreover, $v \Rightarrow_I w$ and, thus, $w \in L^\epsilon_I$. □

From the latter proposition and Theorem 4.2, one derives:

Corollary 4.11. *If $I = \{a^h b\}$ or $I = \{ab^h\}$, with $h \geq 1$, then $\vdash_I^*$ is a wqo on $L^\epsilon_{\vdash_I}$.*

Concluding remark. There are two other simple cases: $I = \{aabb\}$, $I = \{aba\}$. By Theorem 4.4, in the first case $L^\epsilon_{\vdash_I}$ is context-free while in the second one it is not. In both cases we were not able to decide the wqo property of $\vdash_I^*$ on $L^\epsilon_{\vdash_I}$. Thus another interesting problem seems to be the following one: given a finite set I decide whether $\vdash_I^*$ is a wqo on $L^\epsilon_{\vdash_I}$.

Acknowledgements. We kindly acknowledge Gwénaël Richomme for several useful comments on the first version of this paper.

References

[1] D.P. Bovet and S. Varricchio, On the regularity of languages on a binary alphabet generated by copying systems. *Inform. Process. Lett.* **44** (1992) 119–123.

[2] F. D'Alessandro and S. Varricchio, On well quasi-orders on languages. *Lect. Notes Comput. Sci.* **2710** (2003) 230–241,

[3] F. D'Alessandro and S. Varricchio, Well quasi-orders and context-free grammars. *Theor. Comput. Sci.* **327** (2004) 255–268.

[4] A. de Luca and S. Varricchio, Some regularity conditions based on well quasi-orders. *Lect. Notes Comput. Sci.* **583** (1992) 356–371.

[5] A. de Luca and S. Varricchio, Well quasi-orders and regular languages. *Acta Inform.* **31** (1994) 539–557.

[6] A. de Luca and S. Varricchio, *Finiteness and regularity in semigroups and formal languages.* EATCS Monographs on Theoretical Computer Science, Springer, Berlin (1999).

[7] A. Ehrenfeucht, D. Haussler and G. Rozenberg, On regularity of context-free languages. *Theor. Comput. Sci.* **27** (1983) 311–332.

[8] T. Harju and L. Ilie, On well quasi orders of words and the confluence property. *Theor. Comput. Sci.* **200** (1998) 205–224.

[9] D. Haussler, Another generalization of Higman's well quasi-order result on Σ^*. *Discrete Math.* **57** (1985) 237–243.

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

[11] L. Ilie and A. Salomaa, On well quasi orders of free monoids. *Theor. Comput. Sci.* **204** (1998) 131–152.
[12] B. Intrigila and S. Varricchio, On the generalization of Higman and Kruskal's theorems to regular languages and rational trees. *Acta Inform.* **36** (2000) 817–835.
[13] M. Ito, L. Kari and G. Thierrin, Shuffle and scattered deletion closure of languages. *Theor. Comput. Sci.* **245** (2000) 115–133.
[14] M. Jantzen, Extending regular expressions with iterated shuffle. *Theor. Comput. Sci.* **38** (1985) 223–247.
[15] J. Kruskal, The theory of well quasi-ordering: a frequently discovered concept. *J. Combin. Theory Ser. A* **13** (1972) 297–305.
[16] L. Puel, Using unavoidable sets of trees to generalize Kruskal's theorem. *J. Symbolic Comput.* **8** (1989) 335–382.