Well Quasi-orders in Formal Language Theory*

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. The concept of well quasi-order is a generalization of the classical notion of well order and plays a role in the studying of several problems of Mathematics and Theoretical Computer Science. This paper concerns some applications of well quasi-orders to Formal Language Theory. In particular, we present a survey of classical and recent results, based upon such structures, concerning context-free and regular languages. We also focus our attention to some application of well quasi-orders in the studying of languages obtained by using the operators of shuffle and iterated shuffle of finite languages.

Keywords: Well quasi-orders, finite automata, context-free languages, shuffle, iterated shuffle.

1 Introduction

The concept of well quasi-order is a generalization of the classical notion of well order. 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. There exist various characterizations of this concept which was often rediscovered by different authors (see [20]). The concept of well quasi-order plays a role in the studying of many problems of Mathematics and Theoretical Computer Science. For this reason, well quasi-orders have been widely investigated in the past and there exists a large literature on this subject. Recently, in the theory of language equations, remarkable results based on wqo's have been obtained by M. Kunc [22]. These results have been culminating in the negative solution of the famous conjecture by Conway claiming the regularity of the maximal solutions of the *commutative language equation* $XL = LX$ where L is a finite language of words [21]. On the other hand, using wqo's, in [22] it is proved that the maximal solution of the inequality $XK \subseteq LX$ is a regular language whenever L is so. In this paper, we offer a survey of some classical and recent results about the applications of well

* The first author acknowledges the partial support of ``fundings ``Facoltà di Scienze MM. FF. NN. 2006'' of the University of Rome ``La Sapienza''.

M. Ito and M. Toyama (Eds.): DLT 2008, LNCS 5257, pp. 84–95, 2008.

quasi-orders in Formal Language Theory. The first part of the paper presents two basic theorems that give a deep insight into combinatorics on words and languages. The first is due to Higman [15] and it 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.* The second is a remarkable generalization of the famous Myhill-Nerode theorem on regular languages. In [11] Ehrenfeucht et al. proved that a language is regular if and only if it is upwards closed with respect to a monotone well quasi-order. From this result many regularity conditions have been derived (see for instance [1,8,9,10]). Monotone quasi-orders can be associated naturally with the derivation relations of suitable semi-Thue systems, so that one can prove the regularity of a language generated by a semi-Thue system by showing the wqo property of the corresponding derivation relation. In [11] a class of semi-Thue systems called *unitary* is studied. In particular unitary systems whose derivation relation is a wqo are characterized. By applying this result and the generalized Myhill-Nerode theorem, one can obtain a remarkable condition that assures that a language generated by a unitary system is regular. Another important application is the regularity of the languages on a binary alphabet generated by copying systems [1]. We also present a new generalization [2,3] of Higman's theorem to context-free languages while in Section 5 an improvement [4,5] of the above mentioned result for unitary systems is described. In the last section, we consider some applications of well quasi-orders in the studying of languages obtained by using the operators of shuffle and iterated shuffle of finite languages.

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 $w = a_1a_2 \cdots a_n$, 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. 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 $u = a_1a_2 \cdots a_n$, $v = v_1a_1v_2a_2 \cdots v_na_nv_{n+1}$ 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 $L \subseteq B^*$. Moreover, $\mathrm{Pref}(L)$ denotes the set of the prefixes of all words of L. A subset X of a semigroup S is called *recognizable* if there exists a finite index congruence of S that saturates X, that is X is a union of cosets of the congruence. The family of recognizable languages of S is denoted $Rec(S)$. Let P be a subset of a semigroup S. Then the sets $P^{-1}S$ and SP^{-1} are defined as:

$$P^{-1}S = \{t \in S \mid \exists\, p \in P,\ s \in S \mid s = pt\},$$

and

$$SP^{-1} = \{t \in S \mid \exists\, p \in P,\ s \in S \mid s = tp\}.$$

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. [10]).

Theorem 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 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.

Let $\leq$ be a quasi-order on a set S and let X be a subset of S. We say that X is *upwards closed*, or simply *closed*, with respect to $\leq$, if $x \leq y$ and $x \in S$ implies $y \in S$.

Following [10], 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 $p \to q$. 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^*.

3 Generalized Myhill-Nerode Theorem and Highman Theorem

According to the classical Myhill-Nerode theorem, one can obtain a characterization of recognizable subsets of a semigroup in terms of finite index congruence of the semigroup. A remarkable extension of this theorem was obtained in [11] in terms of wqo.

Theorem 2. *A subset X of a semigroup S is recognizable if and only if X is closed with respect to a monotone well quasi-order in S.*

In Sections 4 and 5 we will consider some applications of the previous theorem to formal languages. It is useful to recall that recognizable sets of a semigroup can be described also by using equivalence relations, monotone on the right or on the left. More precisely, a classical theorem by Nerode states that a subset X is recognizable if and only if there exists a finite index equivalence, monotone on the right (resp. on the left) that saturates X. In this context, a result connected with Theorem 2 was proposed in [9,10]. If X is a subset of a semigroup S, then we associate with X a quasi-order $\leq_X^r$ defined as: for any $s, t \in S$,

$$s \leq_X^r t \iff s^{-1}X \subseteq t^{-1}X.$$

The relation $\leq_X^r$ is monotone on the right. Similarly, one can associate a quasi-order, monotone on the left $\leq_X^l$ defined as: $s \leq_X^l t \iff Xs^{-1} \subseteq Xt^{-1}$. In analogy with the theorem by Nerode, one can ask whether the wqo property of $\leq_X^r$ implies the regularity of X. The answer is negative. Indeed, for instance, one can check that the language $L = \{a^n b^m \mid n \geq m \geq 0\}$ is not regular while, on the other hand, $\leq_X^l$ is a wqo. However, a partial generalization of Nerode's theorem and of Theorem 2 as well is the following.

Theorem 3. *A subset X of a semigroup S is recognizable if and only if the quasi-orders $\leq_X^r$ and $\leq_X^l$ are wqo.*

As a consequence one has:

Corollary 1. *A subset X of a semigroup S is recognizable if and only if X is closed with respect to a left and to a right monotone well quasi-order in S.*

Another important result proved in the wqo theory is the Higman theorem. We recall that a quasi-order $\leq$ in a semigroup S is said to be a *division order* or a *divisibility order* if it is monotone and, moreover, for all $s \in S$ and $x, y \in S^1$, $s \leq xsy$. The ordering by divisibility in abstract algebras was studied by Higman who proved in [15] a very general theorem that, in the case of semigroups, has the following statement.

Theorem 4. *Let S be a semigroup quasi-ordered by a divisibility order $\leq$. If there exists a generating set of S well quasi-ordered by $\leq$, then S will be also so.*

It is worth recalling that in [20] Kruskal extends Higman's result, proving that certain embeddings on finite trees are well quasi-orders. Moreover, in [17] some extensions of Higman and Kruskal's theorem to regular languages and rational trees have been given. In particular, we recall the following generalization of Kruskal's theorem:

Theorem 5. *Let A be a wqo alphabet and let T be the family of the rational k-ary trees, with nodes labeled by A. Then the natural embedding relation induced on T is a wqo.*

A remarkable consequence of Theorem 4 is the following. Let $S = A^*$ be the free monoid generated by an alphabet A quasi-ordered by a relation $\leq$. The relation $\leq$ can be extended to A^* as follows. Let $u, v \in A^*$. We set $u \leq v$ if

$$u = a_1 \cdots a_n, a_i \in A,\ i = 1, ..., n,$$

$$v \in A^* b_1 A^* b_2 A^* \cdots A^* b_n A^*, b_i \in A,\ i = 1, ..., n,$$

where

$$a_i \leq b_i, i = 1, ..., n.$$

Trivially, the relation defined above is a division order, called *subsequence ordering*, and if $\leq$ is a wqo on A, then, by Higman theorem, its extension is a wqo on A^*. In the sequel, we refer to this result as the Higman theorem in the free monoid. It can be proved that the subsequence ordering is the smallest division order in A^*.

In [3] a new generalization of Higman theorem has been given. This result is based upon the notion of *division order* on a language: given a language L over the alphabet A, 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^*. Let $G = (V, A, P)$ be a context-free grammar, where $V = \{A_1, \ldots, A_k\}$ is the alphabet of the variables, A is the alphabet of the terminal symbols and P is the set of the productions. For any i, $1 \leq i \leq k$, denote L_i the language generated by G assuming the variable A_i as start symbol. The following theorem holds [3].

Theorem 6. *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 $\leq$ is a division order on L, then $\leq$ is a well quasi-order on L.*

As an immediate corollary of the previous theorem, we have that 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 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. It is possible to give a slight generalization of the notion of division order on languages as follows.

Definition 1. *Let $L \subseteq A^*$ be a language and let $\leq$ be a monotone quasi-order. Then $\leq$ is a* weak division order on L *if for any $u, x, y \in A^*$ such that $u, xuy, xy \in L$, one has $u \leq xuy$.*

We observe that any division order on L is a weak division order on L but the converse is false. Moreover, any weak division order on A^* is a division order. By using some combinatorial arguments akin to that used to prove Theorem 6, one can prove the following theorem.

Theorem 7. *Let L be a context-free language containing the empty word and generated by a context-free grammar with only one variable. Then any weak division order on L is a wqo on L.*

4 Copying Systems

In this section we describe how Theorem 2 has been used to prove the regularity of some relevant formal languages. We consider the case of copying systems and languages generated by them, introduced in [14]. In that paper it is proved that, when the alphabet has cardinality at least three, such languages are not, in general, regular (see Theorem 10). In the case of a binary alphabet, the languages generated by copying systems are actually all regular [1].

Let $A = \{a, b\}$ and let (A, π) be the rewriting system with $\pi = \{(x, xx) \mid x \in A^*\}$. The derivation relation $\Rightarrow_\pi^*$ is called *copying relation.* We can also consider a restricted copying relation denoted as $\Rightarrow_{\pi'}^*$ where

$$\pi' = \{(a, aa), (b, bb), (ab, abab), (ba, baba)\}.$$

Trivially $\Rightarrow_{\pi'}^* \subseteq \Rightarrow_\pi^*$.

Theorem 8. *The derivation relation $\Rightarrow_{\pi'}^*$ is a well quasi-order on A^*.*

Let us remark that one can easily prove (*cf* [1]) that the rewriting system π' is, in fact, equivalent to π. Moreover, π' is the smallest set of rules among those which are equivalent to π. Therefore, the following result easily follows.

Theorem 9. *The derivation relation $\Rightarrow_\pi^*$ is a well quasi-order on A^*.*

Corollary 2. *Let $L \subseteq A^*$ be a language which is closed with respect to $\Rightarrow_\pi^*$. Then L is a regular language.*

Proof. The statement is a consequence of Theorem 9 and Theorem 2.

Let us now consider a free monoid B^* and the copying relation $\Rightarrow_\pi^*$ in B^*. For any $w \in B^*$ we consider the set $L_{w,\pi}$ defined as

$$L_{w,\pi} = \{u \in B^* \mid w \Rightarrow_\pi^* u\}.$$

If a word w contains at least three distinct letters, then the language $L_{w,\pi}$ is not regular [14].

Theorem 10. *Let $w \in B^*$ be a word such that* $\mathrm{Card}(\mathrm{Alph}(w)) \geq 3$*. Then $L_{w,\pi}$ is not regular.*

Proposition 1. *Let B be a finite alphabet and $w \in B^*$. Then $L_{w,\pi}$ is regular if and only if w contains at most two distinct letters.*

Proof. By Theorem 10, if w is a word containing at least three distinct letters, then $L_{w,\pi}$ is not a regular language. Hence, if $L_{w,\pi}$ is regular, then $\mathrm{Card}(\mathrm{Alph}(w)) \leq 2$. Conversely, suppose that $d = \mathrm{Card}(\mathrm{Alph}(w)) \leq 2$. If $d = 0$, then $w = \epsilon$ and $L_{w,\pi} = \{\epsilon\}$ is regular. If $d = 1$, then $w \in a^*$ with $a \in B$ and $L_{w,\pi} = a^{|w|}a^*$ is regular. If $d = 2$, since $L_{w,\pi}$ is closed with respect to $\Rightarrow_\pi^*$, then from Corollary 2 the result follows.

5 Well Quasi-orders and Unitary Grammars

Other applications of great interest of wqo to Formal Language Theory are based upon the notion of *unitary grammar* introduced in [11]. Let us present these results. 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 is denoted by $\Rightarrow_I^*$ (or, simply, $\Rightarrow^*$). We set $L_I^\epsilon = \{u \in A^* \ \mid \ \epsilon \Rightarrow^* u\}$. A language L is called *unitary* if there exists a finite set of words I such that $L = L_I^\epsilon$. Unitary grammars have been introduced in order to study the relationships between the classes of context-free and regular languages. Let us consider this aspect with more attention. Unitary languages are context-free since, given a language L_I^ϵ, a context-free grammar generating L_I^ϵ can be constructed from the set I in the obvious way.

Example 1. Let $A = \{a, b\}$ and let $I = \{ab\}$. One can verify that the language L_I^ϵ is the language of the so called *semi Dyck words over A*. We recall that a word u over the alphabet A is said to be a semi-Dyck word if $|u|_a = |u|_b$ and, moreover, for every prefix p of u, $|p|_a \geq |p|_b$. This language is context-free non regular. Similarly, if $I = \{ab, ba\}$, then L_I^ϵ is the language of *Dyck words over A*, that is, of all words u such that $|u|_a = |u|_b$. The very same result holds for every alphabet $A = \{a_1, ..., a_k, b_1, ..., b_k\}$.

By the well-known Chomsky-Schützenberger theorem, every context-free language is the homomorphic image of the intersection of a Dyck language with a regular one. Since the class of regular languages is closed under homomorphism and intersection and since Dyck languages are unitary, these facts indicate that, at least, some unitary languages capture the non regular aspect of a context-free language. This argument eventually lead to investigate the conditions assuring the regularity of a unitary language. In this theoretical setting, an important theorem proven in [11] is based upon the notion of unavoidable set. This notion is classical and well-known in the field of Combinatorics on Words (see [23], Ch. 1). A set I of words is said to be *unavoidable* (on the set $A = \mathrm{Alph}(I)$) if every sufficiently long word over A has a factor that belongs to I. A set is said to be *avoidable* if it is not unavoidable. The next two examples prefigurate an important characterization, given with Theorem 11 below, of unavoidable sets of words in terms of the wqo property of the unitary grammars.

Example 2. Let A^* be the free monoid generated by the alphabet $A = \{a, b\}$. Set $I = \{a, bb\}$. Then the set I is clearly unavoidable since, any word of length at least 2 contains a factor in I. On the other hand, one can check that the derivation relation $\Rightarrow_I^*$ is a wqo on A^*. The same result holds in the case $I = \{aa, ab, ba, bb\}$

Example 3. Let A^* be the free monoid generated by the alphabet $A = \{a, b\}$. Set $I = \{ab\}$. Then the set I is clearly avoidable since, for instance, every power of the letter a avoids I. On the other hand the derivation relation $\Rightarrow_I^*$ is not a wqo on A^*. Indeed, one can verify that the sequence $\{a^n\}_{n\geq 0}$ is bad with respect to the relation $\Rightarrow_I^*$.

Theorem 11. *Let I be a finite set of A^+ and assume that $A = \mathrm{Alph}(I)$. Then the derivation relation $\Rightarrow_I^*$ is a wqo on A^* if and only if the set I is unavoidable.*

The following remark concerns a noteworth application of Theorem 11.

Example 4. Let A^* be the free monoid genereted by the alphabet A. Obviously, A is unavoidable in A^*. Set $I = A$. Then the derivation relation $\Rightarrow_I^*$ is the subsequence ordering on A^*. According to Theorem 11, the derivation relation $\Rightarrow_I^*$ is a wqo on A^*. Thus we obtain Higman Theorem in the free monoid.

A straighforward corollary of Theorem 11 gives a regularity condition for languages generated by unitary grammars.

Corollary 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.

Example 5. Let us consider again Example 1. Since $I = \{ab\}$, the language L_I^ϵ is the language of *semi Dyck words*. This language is context-free non regular and I is avoidable.

Example 6. Let us consider again Example 2. One can easily check that L_I^ϵ is the shuffle of a^* and $\{bb\}^*$, and thus it is regular.

A short comment on the proof of the corollary above. Since the language L_I^ϵ is closed with respect to the derivation relation $\Rightarrow_I^*$, the implication *ii.* $\Rightarrow$ *iii.* is immediately obtained by applying Theorem 2 to L_I^ϵ. On the other hand, one can prove the implication *iii.* $\Rightarrow$ *i.*, by showing that, if I is an avoidable set, then the language L_I^ϵ is not regular. This last task can be done by using a suitable *anti-pumping argument*.

One can ask if, in Corollary 3, the condition *i.* can be replaced by the weaker condition that the relation $\Rightarrow_I^*$ is a wqo on L_I^ϵ. The positive answer to this question was given in [4,5], by proving the following Theorem 12.

Theorem 12. *The derivation relation $\Rightarrow_I^*$ is a wqo on A^* if and only if $\Rightarrow_I^*$ is a wqo on L_I^ϵ.*

We mention that another important contribution to the field of formal languages whose proof is based upon Corollary 3 was given by Senizergues in [26]. Here it is proved that every rational subset of a free group is either recognizable or *disjunctive*, that is the syntactic congruence associated with the set is the identical relation. This result can be viewed as an extension of the classical Kleene theorem to rational sets of free groups and gives a positive answer to an open problem raised by Sakarovitch. The reader is referred to [25] for a complete survey on this problem.

6 On Other Well Quasi-orders

One can consider a possible extension of the results presented in the previous section, Theorem 12, Corollary 3 and Theorem 11, with respect to other significant quasi orders. If I is a finite set of words, let us associate with I a binary 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 $a_1 a_2 \cdots a_n \in I$. We set $L_{\vdash_I}^\epsilon = \{w \in A^* \mid \epsilon \vdash_I^* w\}$. In [13], the following theorem has been proved.

Theorem 13. *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 $L_{\vdash_I}^\epsilon$ is regular.*

In [13] it is also proved that I is subsequence unavoidable if and only if, for every $a \in A$, $I \cap \{a\}^+ \neq \emptyset$.

Example 7. Let A^* be the free monoid generated by the alphabet $A = \{a, b\}$. Set $I = \{ab\}$. Then the set I is clearly subsequence avoidable since, for instance, every power of the letter a avoids I. Moreover, it is easily seen that the derivation relation $\vdash_I^*$ is not a wqo on A^*. Indeed, one can verify that the sequence $\{a^n\}_{n\geq 0}$ is bad with respect to the relation $\vdash_I^*$. On the other hand, one can verify that $L_{\vdash_I}^{\epsilon} = L_I^{\epsilon}$, so that this language is equal to the language of the semi Dyck words.

Another interesting property of the relation $\vdash_I^*$ is the following consequence of Theorem 7 proven in [3].

Proposition 2. *Let $I \subseteq A^+$. Then $\vdash_I^*$ is a well quasi order on L_I^{ϵ}.*

Proof. The language L_I^{ϵ} is generated by a context-free grammar with only one variable and $\epsilon \in L_I^{\epsilon}$. Moreover, the relation $\vdash_I^*$ is a weak division order over L_I^{ϵ}. The statement, then, follows from Theorem 7.

By the previous proposition, it is natural to ask whether $\vdash_I^*$ is a wqo on $L_{\vdash_I}^{\epsilon}$ or not. The answer is negative. In fact, we can exhibit a set I such that the quasi-order $\vdash_I^*$ is not a wqo on $L_{\vdash_I}^{\epsilon}$. 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 3. *The sequence $\{S_n\}_{n\geq 1}$ is bad with respect to $\vdash_I^*$. Moreover, the elements of $\{S_n\}_{n\geq 1}$ belong to $L_{\vdash_I}^{\epsilon}$ and so $\vdash_I^*$ is not a wqo on $L_{\vdash_I}^{\epsilon}$.*

We can summarize the relationships between, on one hand, the quasi-orders $\vdash_I^*$ and $\Rightarrow_I^*$, and, on the other hand, the languages $L_{\vdash_I}^{\epsilon}$ and L_I^{ϵ} by the following list:

- 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}.

The theoretical setting we have described, suggests to ask whether Theorem 13 may be extended by replacing condition (i) with the weaker condition that the derivation relation $\vdash_I^*$ is a wqo on $L_{\vdash_I}^{\epsilon}$. Unfortunately this is not true. Indeed, by the previous Example 7, if $I = \{ab\}$, $L_{\vdash_I}^{\epsilon} = L_I^{\epsilon}$ is the language of all *semi-Dyck words* over the alphabet $\{a, b\}$. By Proposition 2, $\vdash_I^*$ is a well quasi order on $L_{\vdash_I}^{\epsilon} = L_I^{\epsilon}$ while this language is not regular. This example lead us to further investigate the relation between $L_{\vdash_I}^{\epsilon}$ and $\vdash_I^*$. The results of this investigation will be presented in the next section.

7 Well Quasi-orders and Shuffle Closure of Finite Languages

Given a set I of word, the set $L^{\epsilon}_{\vdash_I}$ is, actually, the set of all the words obtained by the shuffle of (copies of) words of I. Moreover, the relation $\vdash^*_I$ is a natural partial order over $L^{\epsilon}_{\vdash_I}$. Observe also that for any u, v in $L^{\epsilon}_{\vdash_I}$, $u \vdash^*_I v$ if and only if v is the shuffle of u and another word of $L^{\epsilon}_{\vdash_I}$. In [5], the authors have opened the problem of the characterization of the finite sets I such that $\vdash^*_I$ is a well quasi-order on $L^{\epsilon}_{\vdash_I}$. In this section we present the results of [6,7], where a complete answer is given in the case when I consists of a single word w.

In this context, it is worth noticing that in [5] is proved that $\vdash^*_{\{w\}}$ is not a wqo on $\mathrm{L}^{\epsilon}_{\vdash_{\{w\}}}$ if $w = abc$. A simple argument allows one to extend the result above in the case that $w = a^i b^j c^h$, $i, j, h \geq 1$. By using a simple technical argument, this implies that if a word w contains three distinct letters at least, then $\vdash^*_{\{w\}}$ is not a wqo on $\mathrm{L}^{\epsilon}_{\vdash_{\{w\}}}$. Therefore, in order to characterize the word w such that $\vdash^*_{\{w\}}$ is a wqo on $\mathrm{L}^{\epsilon}_{\vdash_{\{w\}}}$, one can consider only the case when w is a word on the binary alphabet $\{a, b\}$. Let E be the exchange morphism ($E(a) = b$, $E(b) = a$), and let $\tilde{w}$ be the mirror image of w.

Definition 2. *A word w is called* bad *if one of the words w, $\tilde{w}$, $E(w)$ and $E(\tilde{w})$ has a factor of one of the two following forms*

$$a^k b^h \quad \textit{with } k, h \geq 2 \tag{1}$$
$$a^k b a^l b^m \; \textit{with } k > l \geq 1, m \geq 1 \tag{2}$$

A word w is called good *if it is not bad.*

One can prove that a word is good if and only if it is a factor of $(ba^n)^\omega$ or $(ab^n)^\omega$ for some $n \geq 0$. The following result characterizes the set of good words in terms of the wqo property [6,7].

Theorem 14. *Let w be a word over the alphabet $\{a, b\}$. The derivation relation $\vdash^*_{\{w\}}$ is a wqo on $\mathrm{L}^{\epsilon}_{\vdash_{\{w\}}}$ if and only if w is good.*

Corollary 4. *Let w be a word over the alphabet $\{a, b\}$. The derivation relation $\vdash^*_{\{w\}}$ is a wqo on $\mathrm{L}^{\epsilon}_{\vdash_{\{w\}}}$ if and only if w is a factor of $(ba^n)^\omega$ or $(ab^n)^\omega$ for some $n \geq 0$.*

References

1. Bovet, D.P., Varricchio, S.: On the regularity of languages on a binary alphabet generated by copying systems. Information Processing Letters 44, 119–123 (1992)
2. D'Alessandro, F., Varricchio, S.: On well quasi-orders on languages. In: Ésik, Z., Fülöp, Z. (eds.) DLT 2003. LNCS, vol. 2710, pp. 230–241. Springer, Heidelberg (2003)
3. D'Alessandro, F., Varricchio, S.: Well quasi-orders and context-free grammars. Theoretical Computer Science 327(3), 255–268 (2004)

4. D'Alessandro, F., Varricchio, S.: Avoidable sets and well quasi orders. In: Calude, C.S., Calude, E., Dinneen, M.J. (eds.) DLT 2004. LNCS, vol. 3340, pp. 139–150. Springer, Heidelberg (2004)
5. D'Alessandro, F., Varricchio, S.: Well quasi-orders, unavoidable sets, and derivation systems. RAIRO Theoretical Informatics and Applications 40, 407–426 (2006)
6. D'Alessandro, F., Richomme, G., Varricchio, S.: Well quasi orders and the shuffle closure of finite sets. In: H. Ibarra, O., Dang, Z. (eds.) DLT 2006. LNCS, vol. 4036, pp. 260–269. Springer, Heidelberg (2006)
7. D'Alessandro, F., Richomme, G., Varricchio, S.: Well quasi-orders and context-free grammars. Theoretical Computer Science 377(1–3), 73–92 (2007)
8. de Luca, A., Varricchio, S.: Some regularity conditions based on well quasi-orders. In: Simon, I. (ed.) LATIN 1992. LNCS, vol. 583, pp. 356–371. Springer, Heidelberg (1992)
9. de Luca, A., Varricchio, S.: Well quasi-orders and regular languages. Acta Informatica 31, 539–557 (1994)
10. de Luca, A., Varricchio, S.: Finiteness and regularity in semigroups and formal languages. EATCS Monographs on Theoretical Computer Science. Springer, Berlin (1999)
11. Ehrenfeucht, A., Haussler, D., Rozenberg, G.: On regularity of context-free languages. Theoretical Computer Science 27, 311–332 (1983)
12. Harju, T., Ilie, L.: On well quasi orders of words and the confluence property. Theoretical Computer Science 200, 205–224 (1998)
13. Haussler, D.: Another generalization of Higman's well quasi-order result on Σ^*. Discrete Mathematics 57, 237–243 (1985)
14. Ehrenfeucht, A., Rozenberg, G.: On regularity of languages generated by copying systems. Discrete Applied Mathematics 8, 313–317 (1984)
15. Higman, G.H.: Ordering by divisibility in abstract algebras. Proc. London Math. Soc. 3, 326–336 (1952)
16. Ilie, L., Salomaa, A.: On well quasi orders of free monoids. Theoretical Computer Science 204, 131–152 (1998)
17. Intrigila, B., Varricchio, S.: On the generalization of Higman and Kruskal's theorems to regular languages and rational trees. Acta Informatica 36, 817–835 (2000)
18. Ito, M., Kari, L., Thierrin, G.: Shuffle and scattered deletion closure of languages. Theoretical Computer Science 245(1), 115–133 (2000)
19. Jantzen, M.: Extending regular expressions with iterated shuffle. Theoretical Computer Science 38, 223–247 (1985)
20. Kruskal, J.: The theory of well quasi-ordering: a frequently discovered concept. J. Combin. Theory, Ser. A 13, 297–305 (1972)
21. Kunc, M.: The power of commuting with finite sets of words. In: Diekert, V., Durand, B. (eds.) STACS 2005. LNCS, vol. 3404, pp. 569–580. Springer, Heidelberg (2005)
22. Kunc, M.: Regular solutions of language inequalities and well quasi-orders. Theoretical Computer Science 348(2-3), 277–293 (2005)
23. Lothaire: Algebraic combinatorics on words. In: Encyclopedia of Mathematics and its applications. Cambridge University Press, Cambridge (2002)
24. Puel, L.: Using unavoidable sets of trees to generalize Kruskal's theorem. J. Symbolic Comput. 8(4), 335–382 (1989)
25. Sakarovitch, J.: Éléments de théorie des automates, Vuibert, Paris (2003)
26. Senizergues, G.: On the rational subsets of the free group. Acta Informatica 33(3), 281–296 (1996)