Wheelerness and Complementation

Giuseppa Castiglione[1], Giovanna D'Agostino[2,*], Alberto Policriti[2], Antonio Restivo[1] and Brian Riccardi[3]

[1]Dip. di Matematica e Informatica, Università di Palermo, Italy
[2]Dip. di Scienze Matematiche, Informatiche e Fisiche, Università degli Studi di Udine, Italy
[3]Dip. di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Italy

Abstract

Wheeler languages, introduced to capture a class of regular languages compatible with an ordered and indexable structure, form a well-behaved subclass of the regular languages. In this paper, we study a little-explored property of such languages: closure under complementation. Specifically, we provide a complete characterization of Wheeler languages whose complement is also Wheeler. Our results offer a deeper understanding of the internal structure of these languages and have both theoretical implications—within the classification of regular languages—and practical applications, particularly in fields leveraging coherent orderings, such as text indexing and genomic data analysis.

Keywords

String Matching, Deterministic Finite Automata, Wheeler Languages, Graph Indexing, Co-lexicographical Sorting

1. Introduction

In recent years, interest in *Wheeler languages*, introduced in [1], has grown significantly, in part due to their central role in efficient indexable data structures such as Wheeler graphs and the *Burrows-Wheeler Transform* (BWT). A regular language is said to be Wheeler if it can be recognized by a deterministic automaton equipped with a total order on its states—i.e. *sets* of strings—that satisfy monotonicity conditions coordinating such ordering with the ordering of its transition labels.

While several studies have investigated the structure, minimization, and verification of the Wheeler property in regular languages (see [2, 3, 4]), and generalize it to a larger context (see [5]), the question of closure under complementation remains largely open. It is known that the class of Wheeler languages (a subclass of star-free languages) is not closed under complement in general, which naturally raises the following questions, first addressed in [6]: which Wheeler languages have the property that their complement is also Wheeler? Can we provide a *structural* characterization of this subclass?

In this work, we answer these questions affirmatively by presenting a formal and constructive characterization of Wheeler languages that are closed under complement within the Wheeler class. Our analysis relies on a combination of automata-theoretic techniques and combinatorial properties of co-lexicographical orders, extending and strengthening existing results. In addition to its theoretical relevance, our characterization helps to delineate more precisely the computational and expressive boundaries of Wheeler-based data structures.

More precisely, our result builds on a characterization given in [7] of the languages $\mathcal{L}$ for which both $\mathcal{L}$ and its complement $\overline{\mathcal{L}}$ are Wheeler with respect to *every* total order on the alphabet. This universality condition is particularly strong and, as a consequence, the class of such languages is rather limited. In fact, it coincides the union of definite (DEF) and reverse definite languages (RDEF). The classes of DEF and RDEF are classical subclasses of regular languages that have been studied since the origins of automata theory [8] and play a role in classifications of star-free languages [9]. Starting from this characterization, our goal is to extend the class DEF $\cup$ RDEF so as to capture exactly those

ICTCS 2025, 26th Italian Conference on Theoretical Computer Science, September 10–12, 2025, Pescara, Italy
*Corresponding author.

languages that are Wheeler and whose complement is also Wheeler, but with respect to a fixed alphabet order, rather than universally. To this end, we treat the classes DEF and RDEF separately. For the reverse definite languages, we introduce a new operation—parameterized by a fixed total order on the alphabet—that, when applied to RDEF, yields a larger class of languages which are guaranteed to be Wheeler and to have a Wheeler complement for that specific order. For definite languages, we first provide a novel characterization of the class DEF in terms of string intervals. We then extend this notion by allowing infinite periodic strings as interval endpoints. As in the RDEF case, this generalized framework allows us to define new languages that are Wheeler and have Wheeler complements for the chosen alphabet order.

The paper is organized as follows. In Section 2 we recall the basic notions regarding strings and regular languages that will be needed to prove our result (see [10] as a general reference). In Section 3 we will summarize all results on Wheeler automata and Wheeler languages together with some particular constructions on DFA that will be needed in the sequel. Section 4 contains the characterization of the class of regular languages which are Wheeler with a Wheeler complement and, finally, Section 5 contains conclusions and open problems.

2. Basics

Strings and Automata. Letting Σ be a finite alphabet, we denote by Σ^* the set of finite words (or strings) over Σ. A *language* $\mathcal{L}$ is a subset of Σ^* and $\mathrm{Pref}(\mathcal{L})$ is the set of all prefixes of words in $\mathcal{L}$. A *deterministic finite automaton* (DFA) is a tuple $\mathcal{D} = (Q, s, \Sigma, \delta, F)$, where Q is a finite non-empty set of *states*, s is the *initial state* (or *source*), the alphabet is Σ, $\delta\colon Q \times \Sigma \to Q$ is the (possibly partial) *transition function*, and $F \subseteq Q$ is the set of *final states*. Whenever δ is not defined on (q, σ) we write $\delta(q, \sigma) = \perp$. If δ is a total function we say that the DFA is *complete*. When clear from the context, we omit the alphabet Σ. If we are dealing with a single DFA, the symbols Q, s, δ, F always refer to it. Sometimes we shall describe the transition function δ using triples (labeled edges), where (q, a, q') stands for $q' = \delta(q, a)$. We call (q, a, q') an *a-transition* and (q, a, q) an *a-loop*. As customary, we extend δ to strings: for all $q \in Q$, $a \in \Sigma$ and $\alpha \in \Sigma^*$, we put $\delta(q, \epsilon) = q$ and $\delta(q, \alpha a) = \delta(\delta(q, \alpha), a)$, where $\delta(\perp, a) = \perp$. Given $\mathcal{D} = (Q, s, \Sigma, \delta, F)$ and $q \in Q$, we denote by I_q the set $\{\, \alpha \in \Sigma^* : \delta(s, \alpha) = q \,\}$. The language *accepted* (or *recognized*) by $\mathcal{D}$ is then defined as $\mathcal{L}(\mathcal{D}) = \bigcup_{q \in F} I_q$. Moreover, for any $q \in Q$, we denote by $\lambda(q)$ the set of all letters $a \in \Sigma$ that label transitions reaching q, i.e., $\lambda(q) :- \{\, a \in \Sigma : (\exists p \in Q)(\delta(p, a) = q) \,\}$. A Σ-loop in a DFA is a state q such that $\delta(q, a) = q$, for all $a \in \Sigma$. By saying that we add a Σ-loop to $\mathcal{A}$ we mean that we add a new state q together with all transitions $\delta(q, a) = q$ for $a \in \Sigma$ (and eventually some other transitions ending in q). A cycle labeled $\gamma = a_1 \dots a_{n-1} \in \Sigma^*$ in a DFA is a sequence of states $q_1, q_2, \dots, q_n = q_1$ such that $\delta(q_i, a_i) = q_{i+1}$, for all $i < n$, and it is simple if $q_1, \dots, q_{n-1}$ are pairwise distinct (in particular, a loop is a simple cycle).

We mostly consider *trimmed* DFAs, that is, automata in which every state is *reachable* (from the initial state) and *useful* (can reach at least one final state). This is not restrictive: every automaton can be turned into a trimmed and equivalent one by simply deleting unreachable states as well as states not reaching at least one final state. In a trimmed automata $\mathcal{D}$ there can be at most one state without incoming transitions, namely s, and every string that can be read starting from s belongs to the set of prefixes, $\mathrm{Pref}(\mathcal{L}(\mathcal{D}))$.

Languages accepted by DFAs form the class of *regular languages*. Given a regular language $\mathcal{L}$, there exists a unique (up to isomorphism) state-wise *minimum* complete DFA accepting $\mathcal{L}$. The states of this minimum DFA correspond to the classes of the Myhill-Nerode equivalence relation $\equiv_{\mathcal{L}}$ on Σ^*, defined as follows (see [10]):

$$\alpha \equiv_{\mathcal{L}} \beta \iff \{\, \gamma \in \Sigma^* : \alpha\gamma \in \mathcal{L} \,\} = \{\, \gamma \in \Sigma^* : \beta\gamma \in \mathcal{L} \,\}.$$

Denoting by $\mathcal{D}^c_{\mathcal{L}}$ the minimum *complete* DFA having as states the classes $[\alpha]_{\mathcal{L}}$ of the Myhill-Nerode equivalence, we have: $s = [\epsilon]_{\mathcal{L}}$, $\delta([\alpha]_{\mathcal{L}}, a) = [\alpha \cdot a]_{\mathcal{L}}$, and $F = \{\, [\alpha]_{\mathcal{L}} \,:\, \alpha \in \mathcal{L} \,\}$. Letting $\mathcal{D}_{\mathcal{L}}$ be the DFA obtained from $\mathcal{D}^c_{\mathcal{L}}$ by considering only classes $[\alpha]_{\mathcal{L}}$ with $\alpha \in \mathrm{Pref}(\mathcal{L})$ and the transitions among

them, one can easily check that $\mathcal{D}_\mathcal{L}$ recognizes $\mathcal{L}$, it is trimmed, and can differ from $\mathcal{D}^c_\mathcal{L}$ for (at most) one non-final Σ-loop. $\mathcal{D}_\mathcal{L}$ is the DFA with minimum number of states among (not necessarily complete) DFAs accepting $\mathcal{L}$.

Given a finite set $S \subseteq \Sigma^*$, we define the *Prefix-tree* of S as the DFA having $\mathrm{Pref}(S)$ as states, ϵ as (root and) initial state, and whose transitions are $\delta(s, a) = sa$, for $s, sa \in \mathrm{Pref}(S)$. Every state w of this DFA is reached by the single word w. The Prefix-tree *acceptor* of a finite set S is the Prefix-tree of S having S as collection of final states.

The class DEF consists of languages of the form $F \cup \Sigma^* G$, where $F, G \subseteq \Sigma^*$ are finite. The class RDEF is the class of languages whose reverse is in DEF, that is, languages of the form $F \cup G\ \Sigma^*$, where F, G are finite. Notice that, for RDEF languages, we can always assume $\mathrm{Pref}(F) \cap G = \emptyset$ and G prefix-free (meaning that there no strings $u, v \in G$ such that u is a prefix of v). Moreover, in the minimum trimmed DFA recognizing an infinite language in RDEF the set of states Q contains a single final Σ-loop q and the transitions restricted to $Q \setminus \{q\}$ build no cycle.

3. Preliminaries

Since in this paper we will be dealing with a finite alphabet Σ and a fixed total order over it, unless otherwise specified, we will always use $\Sigma = \{1, \ldots, k\}$, for some $k \in \mathbb{N} \setminus \{0\}$, with its natural order. We extend such order co-lexicographically to Σ^*, that is, for $\alpha, \beta \in \Sigma^*$, we have $\alpha \leq \beta$ if and only if either α is a suffix of β (denoted by $\alpha \dashv \beta$), or there exist $\alpha', \beta', \gamma \in \Sigma^*$ and $a, b \in \Sigma$, such that $\alpha = \alpha' a \gamma$ and $\beta = \beta' b \gamma$ and $a < b$. Notice that in the co-lexicographical order every string α has an immediate successor, the string 1α, but there are strings without an immediate predecessor, e.g. the string 2. If α is a string, we denote by $|\alpha|$ its length, and by $\alpha[i]$ its i-th character from the left, if $1 \leq i \leq |\alpha|$.

Wheeler Automata and Languages. Wheeler Automata are a special class of DFAs that leverage an a priori fixed order of the alphabet in order to achieve, among other things, efficient compression and indexing (see [1]). Specifically, the co-lexicographic order is lifted from strings to states of the automaton in such a way that p precedes q if and only if, for every α reaching p and β reaching q, α precedes β co-lexicographically. Axioms (W1) and (W2) below do the job.

Definition 3.1 (Wheeler Automaton [1])**.** A Wheeler DFA (WDFA for brevity) is a trimmed DFA endowed with a total order $(Q, \leq)$ on the set of states such that the initial state s has no incoming transitions, it is minimum for $\leq$, and the following two *Wheeler axioms* are satisfied. Let $p' = \delta(p, i)$ and $q' = \delta(q, j)$:

(W1) if $i < j$, then $p' < q'$;
(W2) if $i = j$, $p < q$, and $p' \neq q'$, then $p' < q'$.

Notice that we use the same symbol for the order on the alphabet and the states. Moreover, notice that (W1) implies that a WDFA is *input consistent*, that is, $|\lambda(p)| = 1$ for all states $p \neq s$. Any DFA can be transformed into an equivalent input consistent DFA in $\mathcal{O}\big(|Q| \cdot |\Sigma|\big)$ time by simply creating, for each state $q \in Q$, at most $|\Sigma|$ copies of q, one for each different incoming label of q. An input consistent DFA can also be pictured as a graph where the labels of the edges are moved to the respective target states (a state is labeled i when the transitions reaching q are i-transitions). The initial state, reached by no transitions, is labeled with $\#$, where $\# < i$ for all $i \in \Sigma$. An example of a Wheeler state-labeled DFA is depicted in Figure 1. The unique Wheeler order on the DFA is the one where, for $i, j \in \Sigma$, if $i < j$, then a state labeled i precedes a state labeled j, and the final state labeled 2 (above in the figure) precedes the state labeled 2 below.

It can be proved (see [3, 5]) that, if a total order satisfying Definition 3.1 exists, then it is unique and, as we said, allows the lifting of the co-lexicographic order from strings to states. More precisely, given a trimmed, input consistent DFA $\mathcal{D}$ in which the initial state has no incoming edges, we can define a

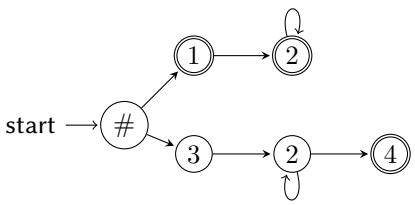

Figure 1: A state-labeled WDFA $\mathcal{D}$ recognizing the language $\mathcal{L} = 12^* \cup 32^+4$.

partial order $\leq_{\mathcal{D}}$ on Q by $q \leq_{\mathcal{D}} q' \Leftrightarrow (q = q') \vee (\forall \alpha \in I_q)(\forall \beta \in I_{q'})(\alpha < \beta)$. Then, it can be proved that $\leq_{\mathcal{D}}$ always satisfies properties (W1) and (W2) and the following holds.

Lemma 1 ([5]). *A trimmed, input consistent DFA is Wheeler if and only if the partial order $\leq_{\mathcal{D}}$ is total.*

Remark 1. Since in a WDFA the order $\leq_{\mathcal{D}}$ is total, we can decide whether $q \leq_{\mathcal{D}} q'$ by simply checking the relative co-lexicographical order among a pair (α, β) with $\alpha \in I_q$ and $\beta \in I_{q'}$. Moreover, if a sequence of words $(\alpha_i)_{i \in \mathbb{N}}$ is monotone in the co-lexicographical order of words, the corresponding sequence of states $(\delta(s, \alpha_i))_{i \in \mathbb{N}}$ must be monotone in the order $\leq_{\mathcal{D}}$ and, since there are only a finite number of states, the sequence $(\delta(s, \alpha_i))_{i \in \mathbb{N}}$ must be eventually constant.

A *Wheeler language* is a language accepted by a Wheeler DFA. We denote by $\mathrm{W}(<)$ the class of languages that are Wheeler in the ordered alphabet $(\Sigma, <)$. In order to recognize if a given regular language is Wheeler, we shall use the following.

Lemma 2 ([3]). *A regular language $\mathcal{L}$ is Wheeler if and only if all monotone sequences in $(\mathrm{Pref}(\mathcal{L}), \leq)$ become eventually constant modulo $\equiv_{\mathcal{L}}$. In other words, for all sequences $(\alpha_i)_{i \in \omega}$ in $\mathrm{Pref}(\mathcal{L})$ such that:*

$$\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_i \leq \ldots \quad \textit{or} \quad \alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_i \geq \ldots,$$

there exists an n such that $\alpha_h \equiv_{\mathcal{L}} \alpha_k$, for all $h, k \geq n$.

Notice that there exists a Wheeler language $\mathcal{L}$ for which the minimum DFA $\mathcal{D}_{\mathcal{L}}$ is not Wheeler. Consider e.g. the DFA given in Fig. 2. This DFA is (minimum) but not Wheeler by Lemma 1: the strings $124, 1244$ reach z, the string 324 reaches state r, and $124 < 324 < 1244$, so that states r, z are $<_{\mathcal{D}}$-incomparable. However the language recognized by this DFA is Wheeler. To prove this, we can use a characterization of (non) Wheeler languages based on the existence of "special" pairs of $\mathcal{D}_{\mathcal{L}}$-incomparable states. More precisely:

Definition 3.2. Let $\mathcal{D} = (Q, s, \Sigma, \delta, F)$ be a DFA. If $p, q \in Q$ then we denote by $p \bowtie_< q$ the fact that p, q are $\leq_{\mathcal{D}_{\mathcal{L}}}$-incomparable, that is:

$$p \bowtie_< q \iff \exists \alpha, \alpha' \in I_p,\ \exists \beta, \beta' \in I_q\ (\alpha < \beta) \wedge (\beta' < \alpha').$$

Theorem 3 ([4]). *Let $\mathcal{D}_{\mathcal{L}}$ be the minimum trimmed DFA accepting $\mathcal{L}$. Then, $\mathcal{L} \notin \mathrm{W}(<)$ if and only if in $\mathcal{D}_{\mathcal{L}}$ there are two nodes $p \neq q$ such that:*

1. *p, q are the starting points of two equally labeled cycles;*
2. *$p \bowtie_< q$.*

Going back to the language of Fig. 2, we notice that $\mathcal{D}_{\mathcal{L}}$ contains only one pair from which two equally labeled cycles start, namely the pair (p, q), but $p <_{\mathcal{D}_{\mathcal{L}}} q$ holds. Hence, the language $\mathcal{L}(\mathcal{D})$ is Wheeler by Theorem 3.

Another consequence of Theorem 3 is that Wheeler languages are star-free. To see this, remember that a *counter* in a DFA is a sequence of $n \geq 2$ pairwise distinct states $q_1, \ldots, q_n$ such that there exists a string $\alpha \in \Sigma^*$ for which: (i) $\delta(q_n, \alpha) = q_1$, and (ii) $\delta(q_i, \alpha) = q_{i+1}$ for all $1 \leq i < n$. It is known that

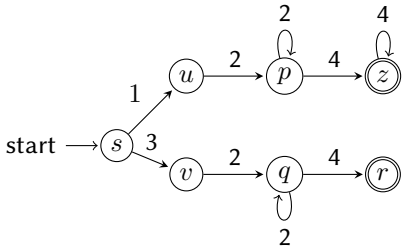

Figure 2: A non Wheeler minimum DFA $\mathcal{D}$ recognizing a Wheeler language.

a language is star-free iff the minimum automaton of the language is counter-free. Hence, Wheeler languages are star-free since one can prove that a counter would always violate the conditions in Theorem 3.

4. Wheeler Languages with a Wheeler complement

Wheeler languages are not closed under complementation; for example, the language 2^* over the ordered alphabet $\{1, 2\}$ is Wheeler but its complement is not: indeed, the monotone sequence

$$2 < 12 < 22 < ... < 2^n < 12^n < 2^{n+1} < ...$$

belongs to $\mathrm{Pref}(\overline{\mathcal{L}})$ but it is not eventually constant modulo $\equiv_{\overline{\mathcal{L}}}$. Hence, $\overline{\mathcal{L}}$ is not Wheeler by Lemma 2.

Remark 2. One can easily find a sufficient condition for describing languages which are Wheeler with a Wheeler complement as follows. Consider a language $\mathcal{L}$ recognized by a *Wheeler complete DFA* $\mathcal{D}$, that is an automaton satisfying Def. 3.1 even without trimming. Clearly, since the trimmed version of $\mathcal{D}$ is still a Wheeler DFA, $\mathcal{L}$ is Wheeler. Moreover, $\overline{\mathcal{L}}$ is also Wheeler, because if we swap final and non final states in $\mathcal{D}$ and trim the resulting DFA, we obtain a Wheeler DFA for $\overline{\mathcal{L}}$. However, this is not a necessary condition for a language to be Wheeler with a Wheeler complement, as the following example shows. Consider $\Sigma = \{1, 2\}$ and the language $1\Sigma^*$. Then $\overline{\mathcal{L}} = \{\epsilon\} \cup 2\Sigma^*$. Using Lemma 2 one can easily check that both $\mathcal{L}$ and $\overline{\mathcal{L}}$ are Wheeler. However, there cannot be a complete Wheeler DFA recognizing any of them. By Remark 2, in such a DFA any monotone sequence of words should eventually end in the same state. However, the words of the sequence

$$2 < 12 < 22 < ... < 2^n < 12^n < 2^{n+1} < ...$$

belong alternatively to $\mathcal{L}$ and $\overline{\mathcal{L}}$, leading to a contradiction. [1]

In [7] it is proved that $\mathrm{DEF} \cup \mathrm{RDEF}$ coincides with the class of languages $\mathcal{L}$ such that both $\mathcal{L}$ and $\overline{\mathcal{L}}$ are Wheeler with respect to *every* order of the alphabet.

Lemma 4 ([7]). *$\mathcal{L}, \overline{\mathcal{L}} \in W(<)$ for all total orders $<$ $\Leftrightarrow$ $\mathcal{L} \in \mathrm{DEF} \cup \mathrm{RDEF}$.*

What happens if we drop the universality condition above? In this paper, we show that in order to characterize the class of languages $\mathcal{L}$ such that $\mathcal{L}, \overline{\mathcal{L}} \in W(<)$ with respect to a *fixed* order $<$ of Σ, it is sufficient to consider a slight extension of the class $\mathrm{DEF} \cup \mathrm{RDEF}$. We start with an example of a language $\mathcal{L}$ such that both $\mathcal{L}$ and its complement $\overline{\mathcal{L}}$ are Wheeler for a *fixed* order, but $\mathcal{L}$ is not in $\mathrm{DEF} \cup \mathrm{RDEF}$.

Example 4.1. Let $\Sigma = \{1, 2\}$ and consider the language $\mathcal{L} = 1^+$ over Σ, recognized by the DFA $\mathcal{D}$ in Fig. 3. This DFA is a *complete* Wheeler DFA, since it satisfies the conditions of Def. 3.1 even without trimming — just consider the order $s <_{\mathcal{D}} q_1 <_{\mathcal{D}} q_2 <_{\mathcal{D}} q_3$. Hence, both $\mathcal{L}$ and $\overline{\mathcal{L}}$ are Wheeler by Remark 2. Moreover, $1^n \in \mathcal{L}$, while $21^n, 1^n2 \notin \mathcal{L}$, for all $n > 1$: hence, we cannot decide membership

[1] Notice that this sequence is neither in $\mathrm{Pref}(\mathcal{L})$ nor in $\mathrm{Pref}(\overline{\mathcal{L}})$, hence it does not contradict the Wheelerness of $\mathcal{L}$ or $\overline{\mathcal{L}}$.

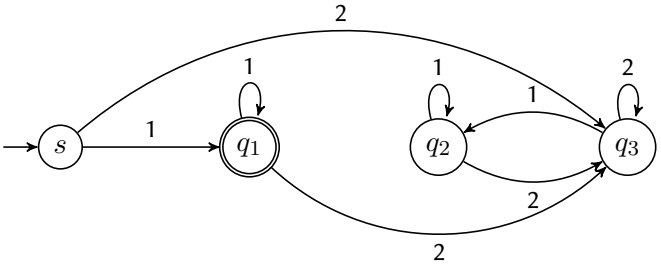

Figure 3: The language recognized by this DFA and its complement are both Wheeler but outside the class $\mathrm{DEF} \cup \mathrm{RDEF}$.

in $\mathcal{L}$ by just checking a prefix or suffix of fixed length, proving that $\mathcal{L} \notin \mathrm{DEF} \cup \mathrm{RDEF}$.
This simple example can also be used to explain another peculiarity of complementation for Wheeler language. If we change the order of the alphabet so that $2 < 1$, then we can prove that the language 1^+ is still Wheeler but its complement is not. Consider again the complete DFA $\mathcal{D}$ of Fig. 3: if $2 < 1$ states q_1, q_2 are incomparable with respect to $<_{\mathcal{D}}$, since $1 < 121 < 11$ and $1, 11$ reach q_1 while 121 reaches q_2. Hence, this DFA is not Wheeler w.r.t. the order $2 < 1$ and we cannot count on it to prove that 1^+ is Wheeler. However, if we trim $\mathcal{D}$ we obtain a Wheeler DFA for 1^+ (insensitive to the order of the alphabet), thereby proving that 1^+ is still Wheeler when $2 < 1$. Notice that this trimmed DFA is useless to show the Wheelerness of the complement, which can indeed proved to be a non Wheeler language by Lemma 2, since the sequence $1 < 121 < 11 < 1211 < 111 < 1211 < \dots$ is monotone in $(\mathrm{Pref}(\overline{\mathcal{L}}), <)$ but it is not eventually constant modulo $\equiv_{\overline{\mathcal{L}}}$. [2]

When the order of the alphabet is fixed, a special role is played, as we shall see, by the first symbol of the alphabet. This can be explained as follows. If we fix a symbol i of the alphabet, the concatenation to the left with i, that is, the function from strings to strings defined by $\alpha \mapsto i\alpha$ is not in general monotone: e.g. $1 < 21$ but $31 > 321$. However, if $i = 1$ is the minimum of $(\Sigma, <)$, then, for any $\alpha \in \Sigma^*$, 1α is the *immediate successor* of α and, therefore, for $\alpha < \beta$ we have:

$$\alpha < 1\alpha \leq \beta < 1\beta.$$

From this it follows, for example, that $\alpha < \beta$ if and only if $1\alpha < 1\beta$. This helps significantly when analyzing Wheelerness of languages.

The following definition characterizes more precisely the kind of application we need from the above considerations.

Definition 4.1. Let $\mathcal{L}$ be a regular language and $\ell = \sup\{\, n \in \mathbb{N} : 1^n \in \mathrm{Pref}(\mathcal{L}) \,\}$. If $\ell = \infty$, then we define $1^{\uparrow}\mathcal{L} = \mathcal{L}$ otherwise, $1^{\uparrow}\mathcal{L} = \mathcal{L} \cup \left\{\, 1^{\ell+n}\beta : 1^{\ell}\beta \in \mathcal{L} \wedge n \in \mathbb{N} \,\right\}$.

Remark 3. If $\ell = 0$, then $1^{\uparrow}\mathcal{L} = 1^*\mathcal{L}$. For example, if $\mathcal{L} = \{212, 3\}$ over the alphabet $\{1, 2, 3\}$, then $1^{\uparrow}\mathcal{L} = 1^*\{212, 3\}$. Notice that the operator $1^{\uparrow}\mathcal{L}$ is idempotent, but it is not a closure operator because it is not monotone, e.g. $\{1, 11\} \subseteq \{1, 11, 1111\}$ but $1^{\uparrow}\{1, 11\} = \{1\} \cup 1^*11 = 1^+ \not\subseteq 1^{\uparrow}\{1, 11, 1111\} = \{1, 11\} \cup 1^*1111$.

We first prove that the operation $\mathcal{L} \mapsto 1^{\uparrow}\mathcal{L}$ preserves Wheelerness. The intuition is that since 1α is the immediate co-lexicographic successor of α, if we add to a language the monotone sequence $1\alpha, 11\alpha, 111\alpha \dots$, with all strings reaching the same state, this cannot create any new monotone sequence violating Lemma 2.

Lemma 5. *If* $\mathcal{L} \in W(<)$ *then* $1^{\uparrow}\mathcal{L} \in W(<)$.

Proof. Let $\mathcal{L} \in W(<)$ and let $\ell = \sup\{\, n \in \mathbb{N} : 1^n \in \mathrm{Pref}(\mathcal{L}) \,\}$. If $\ell = \infty$, then $1^{\uparrow}\mathcal{L} = \mathcal{L}$ and there is nothing to prove, hence assume $\ell < \infty$. Since $\mathcal{L} \in W(<)$, there exists a Wheeler DFA

[2] Notice that this sequence is not in $\mathrm{Pref}(\mathcal{L})$.

$\mathcal{D} = (Q, s, \delta, F)$ recognizing $\mathcal{L}$. We build a Wheeler DFA $\mathcal{D}_1$ recognizing $1^{\uparrow}\mathcal{L}$ in two phases. First, we construct a (Wheeler) $\mathcal{D}' = (Q', s', \delta', F')$ with $\mathcal{L}(\mathcal{D}') = \mathcal{L}$ where, for $k \leq \ell$, any state reached by $\delta'(s', 1^k)$ is *only* reached by 1^k. Then, we turn $\mathcal{D}'$ into a Wheeler DFA $\mathcal{D}_1$ recognizing $1^{\uparrow}\mathcal{L}$, so that $1^{\uparrow}\mathcal{L} = \mathcal{L}(\mathcal{D}_1) \in W(<)$.

The automaton $\mathcal{D}' = (Q', s', \delta', F')$ is defined as follows. For any $1 \leq k \leq \ell$, let q'_k be a copy of the state $q_k = \delta(s, 1^k)$. Let $Q' = Q \cup \{q'_1, \ldots, q'_\ell\}$, $s' = s$, and $F' = F \cup \{\, q'_k : q_k \in F, 1 \leq k \leq \ell \,\}$. The new transition function δ' is obtained from δ by just erasing $(s, 1, q_1)$ and adding $(s, 1, q'_1), (q'_k, 1, q'_{k+1}), (q'_k, j, q)$, for $1 \leq k < \ell, j \neq 1$, and $(q_k, j, q) \in \delta$ (see Fig. 4).

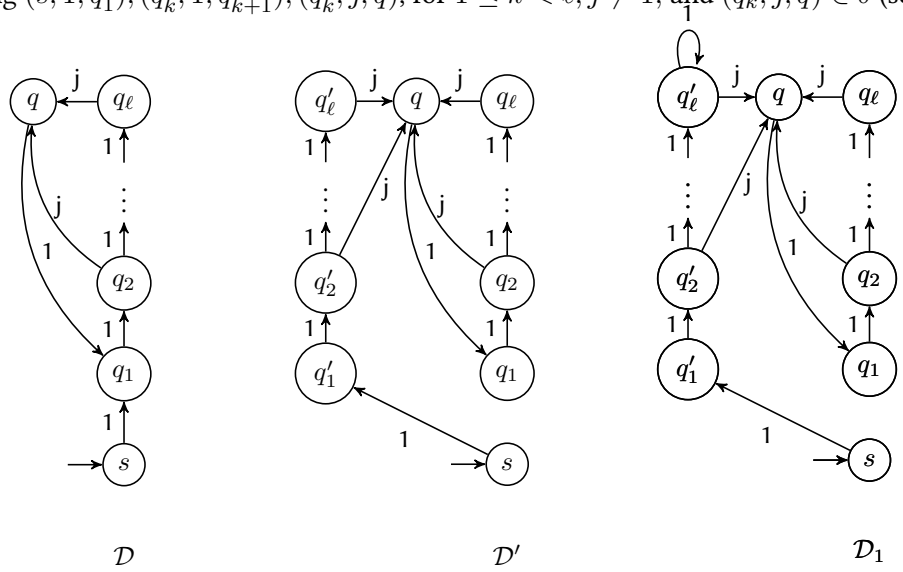

Figure 4: Assuming $j \neq 1$, the picture illustrates the two steps of the construction of $\mathcal{D}_1$ in Lemma 5.

Notice that both $\delta'(q'_\ell, 1)$ and $\delta(q_\ell, 1)$ are undefined. By construction, $\mathcal{D}'$ is a DFA and for any $1 \leq k \leq \ell$ we have that q'_k is reached in $\mathcal{D}'$ by 1^k only. Moreover, words reaching $q \in Q$ in $\mathcal{D}'$ are exactly those reaching q in $\mathcal{D}$ that are different from 1^k, for $1 \leq k \leq \ell$. This implies that $\mathcal{L}(\mathcal{D}') = \mathcal{L}(\mathcal{D}) = \mathcal{L}$. By possibly erasing non-reachable states we can also assume that $\mathcal{D}'$ is trimmed. Moreover, we can check that the preorder $\leq_{\mathcal{D}'}$ is total. Since 1 is the minimum letter of the alphabet, we have $q'_1 <_{\mathcal{D}'} q'_2 <_{\mathcal{D}'} \cdots <_{\mathcal{D}'} q'_\ell$ and all the q'_i's precede states in $Q' \setminus \{q'_1, \ldots, q'_\ell\}$, which are ordered as in $\mathcal{D}$.

From $\mathcal{D}'$ we obtain $\mathcal{D}_1$, recognizing $1^{\uparrow}\mathcal{L}$, by adding the self-loop $(q'_\ell, 1, q'_\ell)$, that is, $\mathcal{D}_1 = (Q', s', \delta' \cup \{(q'_\ell, 1, q'_\ell)\}, F')$ (see Fig. 4). By construction, q'_k, for $1 \leq k < \ell$, is reached in $\mathcal{D}_1$ by 1^k only, while q'_ℓ is reached by the infintely many words in $1^\ell 1^*$. Moreover, a state $q \in Q' \setminus \{q'_1, \ldots, q'_k\} \subseteq Q$ is reached in $\mathcal{D}_1$ only by the words $\alpha \notin 1^*$ and such that one of the following holds:

1. α reaches q in $\mathcal{D}'$ as well, or
2. $\alpha = 1^\ell 1^h \beta$ with $h > 0$, $1^\ell \beta$ reaches q in $\mathcal{D}'$, and $\beta[1] \neq 1$.

From this it follows that also the order $\leq_{\mathcal{D}_1}$ is total. Since, for $1 \leq k < \ell$, q'_k is reached in $\mathcal{D}_1$ only by 1^k and q'_ℓ is reached by the words in $1^\ell 1^*$, we have $q'_1 <_{\mathcal{D}_1} \cdots <_{\mathcal{D}_1} q'_\ell$. Moreover, since 1 is the minimum in Σ, all states in $\{q'_1, \ldots, q'_k\}$ precede states in $Q' \setminus \{q'_1, \ldots, q'_k\}$. Consider now $q, q' \in Q' \setminus \{q'_1, \ldots, q'_k\}$, with $q \neq q'$. Since, as proved above, the order $\leq_{\mathcal{D}'}$ over the automaton $\mathcal{D}'$ is total, q, q' are comparable in $\leq_{\mathcal{D}'}$: say $q <_{\mathcal{D}'} q'$. We now prove that this implies that $q \leq_{\mathcal{D}_1} q'$ holds as well. Suppose α reaches q and α' reaches q' in $\mathcal{D}_1$. Then $\alpha < \alpha'$ because one of the following cases apply:

- α, α' reach q, q' in $\mathcal{D}'$, respectively, hence $\alpha < \alpha'$ because $q <_{\mathcal{D}'} q'$.
- α reaches q in $\mathcal{D}'$, $\alpha' = 1^\ell 1^h \beta$ with $h > 0$, $1^\ell \beta$ reaches q' in $\mathcal{D}'$, and $\beta[1] \neq 1$; in this case, from $q <_{\mathcal{D}'} q'$ it follows $\alpha < 1^\ell \beta < 1^\ell 1^h \beta = \alpha'$.

- $\alpha = 1^\ell 1^h \beta$ with $h > 0$, $1^\ell\beta$ reaches q in $\mathcal{D}'$, and $\beta[1] \neq 1$, while α' reaches q' in $\mathcal{D}'$; in this case, from $q <_{\mathcal{D}'} q'$ it follows $1^\ell\beta < \alpha'$; moreover, all words in $1^\ell 1^+\beta$ reach q in $\mathcal{D}_1$ and form the infinite chain of immediate successors of $1^\ell\beta$ in Σ^*; since $\alpha' \notin 1^\ell 1^+\beta$ (α' reaches q' and not in q in $\mathcal{D}_1$), it follows that $1^\ell 1^h\beta < \alpha'$.
- $\alpha = 1^\ell 1^h \beta$ with $h > 0$, $1^\ell\beta$ reaches q in $\mathcal{D}'$, and $\beta[1] \neq 1$, $\alpha' = 1^\ell 1^k \beta'$ with $k > 0$, $1^\ell\beta'$ reaches q' in $\mathcal{D}'$, and $\beta'[1] \neq 1$. In this case, reasoning as in the previous points we get

$$1^\ell\beta < 1^\ell 1^h\beta = \alpha < 1^\ell\beta' < 1^\ell 1^k\beta' = \alpha'.$$

Hence, we proved that the partial order $\leq_{\mathcal{D}_1}$ is, in fact, total. By Lemma 1 we have that $\mathcal{D}_1$ is a Wheeler automaton and, since it recognizes $1^\uparrow\mathcal{L}$, we have that $1^\uparrow\mathcal{L} \in W(<)$. □

Definition 4.2. The class $1^\uparrow$RDEF is the closure of RDEF under the operator $1^\uparrow\mathcal{L}$:

$$1^\uparrow\mathrm{RDEF} = \mathrm{RDEF} \cup \left\{ 1^\uparrow\mathcal{L} : \mathcal{L} \in \mathrm{RDEF} \right\}$$

We first prove that this class is closed under boolean operations.

Lemma 6. *The class* $1^\uparrow$RDEF *is closed under boolean operations. Moreover,* $1^\uparrow\mathrm{RDEF} \subseteq \mathrm{W}(<)$.

Proof. We first prove closure under complementation. If $\mathcal{L} \in$ RDEF then it is known that $\overline{\mathcal{L}} \in$ RDEF. If $\mathcal{L} \in 1^\uparrow\mathrm{RDEF} \setminus$ RDEF then there exists $\mathcal{R} \in$ RDEF such that $\mathcal{L} = 1^\uparrow\mathcal{R}$ and $1^\uparrow\mathcal{R} \neq \mathcal{R}$, that is $\ell = \sup\{n \ : \ 1^n \in \mathrm{Pref}(\mathcal{R})\} < \infty$. Let F, G be finite sets such that $\mathcal{R} = F \cup G\,\Sigma^*$, $\mathrm{Pref(F)} \cap \mathrm{G} = \emptyset$, and G is prefix-free. Consider the DFA $\mathcal{D}$, recognizing the language $\mathcal{L} = 1^\uparrow\mathcal{R}$, defined as follows. First, we build the Prefix-tree acceptor of $F \cup G$, and since $\mathrm{Pref(F)} \cap \mathrm{G} = \emptyset$, we have that no state in G leads to a state in F. We then add a new Σ-loop $\overline{q}$ and all transitions $(q, i, \overline{q})$, for $q \in G$, obtaining an automaton recognizing $\mathcal{R}$. Finally, we add the self-loop $\delta(1^\ell, 1) = 1^\ell$ obtaining $\mathcal{D}$ recognizing $\mathcal{L} = 1^\uparrow\mathcal{R}$.

In order to obtain a DFA for $\overline{\mathcal{L}}$, starting from $\mathcal{D}$, switch final and non-final states and add a new final Σ-loop q', reached by all missing transitions. Finally, we trim all states not reaching a final state, obtaining a new $\mathcal{D}'$, easily seen to accept $\overline{\mathcal{L}}$. Notice that the Σ-loop $\overline{q}$ of $\mathcal{D}$ will be erased in $\mathcal{D}'$, because it is final in $\mathcal{D}$ and there are no paths leaving it. Hence $\overline{q}$ is not in $\mathcal{D}'$.

We consider two cases.

If state 1^ℓ is erased in $\mathcal{D}'$, then $\mathcal{D}'$ contains only one state reached by infinitely many strings — namely, the Σ-loop q'. In this case $\overline{\mathcal{L}} \in \mathrm{RDEF} \subseteq 1^\uparrow\mathrm{RDEF}$ and we are done.

If 1^ℓ is in $\mathcal{D}'$, then $\mathcal{D}'$ contains only two states reached by infinitely many strings — namely, q' and 1^ℓ —, and the only simple cycles in $\mathcal{D}'$ are the 1-loop on 1^ℓ and the Σ-loop on q'. If we remove the transition $\delta(1^\ell, 1)$ we obtain a DFA $\mathcal{D}''$ recognizing a language $\mathcal{S} \in \mathrm{RDEF}$ such that $\overline{\mathcal{L}} = 1^\uparrow\mathcal{S}$, proving that $\overline{\mathcal{L}} \in 1^\uparrow\mathrm{RDEF}$.

As for closure under union, if $\mathcal{L}_1, \mathcal{L}_2, \in 1^\uparrow\mathrm{RDEF}$, let $\mathcal{L}_1 = 1^\uparrow\mathcal{R}_1$ and $\mathcal{L}_2 = 1^\uparrow\mathcal{R}_2$, with $\mathcal{R}_1, \mathcal{R}_2 \in$ RDEF. Let $\ell_i = \sup\{n \ : \ 1^n \in \mathrm{Pref}(\mathcal{R}_i)\}$, for $i = 1, 2$. If $\ell_1 = \ell_2$ then $1^\uparrow\mathcal{R}_1 \cup 1^\uparrow\mathcal{R}_2 = 1^\uparrow(\mathcal{R}_1 \cup \mathcal{R}_2)$ and we are done. If $\ell_1 < \ell_2$ then consider the language $\tilde{\mathcal{R}}_1 = \mathcal{R}_1 \cup \left\{ 1^h\beta : \ell_1 \leq h \leq \ell_2,\ 1^{\ell_1}\beta \in \mathcal{R}_1 \right\}$. Then $\tilde{\mathcal{R}}_1 \in$ RDEF, because whether a word w belongs to $\tilde{\mathcal{R}}_1$ still depends only on the prefix of fixed length of the word. Hence, $\tilde{\mathcal{R}}_1 \cup \mathcal{R}_2 \in$ RDEF and $\mathcal{L}_1 \cup \mathcal{L}_2 = 1^\uparrow(\tilde{\mathcal{R}}_1 \cup \mathcal{R}_2)$.

Being closed under complementation and union, the class $1^\uparrow$RDEF is closed under boolean operations.

Finally, since $1^\uparrow\mathrm{RDEF} = \mathrm{RDEF} \cup \left\{ 1^\uparrow\mathcal{L} : \mathcal{L} \in \mathrm{RDEF} \right\}$ the inclusion $1^\uparrow\mathrm{RDEF} \subseteq \mathrm{W}(<)$ follows from Lemma 4 and Lemma 5. □

In Lemma 6 we proved that $1^\uparrow\mathrm{RDEF} \subseteq \mathrm{W}(<)$. We obtain a partial converse of this inclusion as follows.

Lemma 7. *Let* $\mathcal{L}$ *be a regular language. If* $\mathrm{Pref}(\overline{\mathcal{L}}) \neq \Sigma^*$ *then*

$$\mathcal{L} \in W(<) \ \Rightarrow \ \mathcal{L} \in 1^\uparrow\mathrm{RDEF}.$$

Proof. Suppose that $\mathcal{L} \in \mathrm{W}(<)$. Since we are assuming $\mathrm{Pref}(\overline{\mathcal{L}}) \neq \Sigma^*$, it follows that in $\mathcal{D}_{\mathcal{L}}$—minimum DFA accepting $\mathcal{L}$—there is a final Σ-loop q. If the only simple cycles in $\mathcal{D}_{\mathcal{L}}$ are the loops over q, then $\mathcal{L} \in \mathrm{RDEF}$ and we are done.

Otherwise, consider a simple cycle *not* visiting q. We claim that if γ is the label of (any) such a simple cycle $\mathcal{C}$ starting from a state p and α is such that $\delta(s, \alpha) = p$, then $\gamma = 1$ and $\alpha = 1^m$ for some $m \in \mathbb{N}$. This implies that such a cycle is unique, namely, a 1-loop on $\delta(s, 1^m)$. We prove our claim showing that, otherwise, we could find three words $\zeta_1 < \zeta_2 < \zeta_3$ with ζ_1, ζ_3 reaching q and ζ_2 reaching p, respectively. This would imply $p \bowtie_< q$ and, since γ is the label of a cycle both from p and from (the Σ-loop) q, by Theorem 3 we would contradict $\mathcal{L} \in \mathrm{W}(<)$.

Let ξ be a word reaching q. We may assume $\xi \neq \epsilon$, otherwise $\mathcal{L} = \Sigma^*$ and we are done. Let $\alpha' \in \Sigma^*$ be any word with $\alpha < \alpha'$. Suppose first that $\gamma \notin 1^*$. Then, there exists $\gamma' < \gamma$ with $|\gamma'| = |\gamma|$ and we can consider $\zeta_1 = \xi\gamma' < \zeta_2 = \alpha\gamma < \zeta_3 = \xi\alpha'\gamma$. Since $\zeta_1, \zeta_3 \in I_q$, and $\zeta_2 \in I_p$, we are done. Hence, $\gamma \in 1^*$ and, furthermore, since $\mathcal{L} \in W(<)$ and, consequently, is star-free, $\mathcal{D}_{\mathcal{L}}$ must be counter-free, implying $\gamma = 1$.

Next, suppose that $\alpha \notin 1^*$. Then let $\alpha', \alpha'' \in \Sigma^*$ be such that $|\alpha''| = |\alpha|$ and $\alpha'' < \alpha < \alpha'$. Since $\zeta_1 = \xi\alpha'', \zeta_2 = \alpha, \zeta_3 = \xi\alpha'$ are such that $\zeta_1 < \zeta_2 < \zeta_3$, we would again contradict $\mathcal{L} \in W(<)$. Therefore, $\alpha \in 1^*$ and the claim is proved.

The only cycle in $\mathcal{D}_{\mathcal{L}}$ not visiting q is the 1-labelled self-loop on p, with p reached by 1^m only. Erasing the 1-loop from p we then obtain a DFA accepting $\mathcal{R} \in \mathrm{RDEF}$, thereby proving $\mathcal{L} = 1^{\uparrow}\mathcal{R} \in 1^{\uparrow}\mathrm{RDEF}$. □

Corollary 8. *If $\mathcal{L}$ is a regular language with $\mathcal{L} \in W(<)$ and $\mathrm{Pref}(\overline{\mathcal{L}}) \neq \Sigma^*$, then $\overline{\mathcal{L}} \in W(<)$.*

Proof. Suppose $\mathcal{L} \in W(<)$ and $\mathrm{Pref}(\overline{\mathcal{L}}) \neq \Sigma^*$. By Lemma 7 we have that $\mathcal{L} \in 1^{\uparrow}\mathrm{RDEF}$, hence $\overline{\mathcal{L}} \in 1^{\uparrow}\mathrm{RDEF} \subseteq \mathrm{W}(<)$ by Lemma 6. □

Corollary 9. *If $\mathcal{L}$ is a regular language and $\mathrm{Pref}(\overline{\mathcal{L}}) \neq \Sigma^*$ then,*

$$\mathcal{L} \in W(<) \Leftrightarrow \mathcal{L} \in 1^{\uparrow}\mathrm{RDEF}.$$

Proof. Implication from left to right is given in Lemma 7. Implication from right to left follows from Lemma 6. □

Thanks to the previous results we now have a complete characterization of the Wheeler Languages with a Wheeler complement in the cases in which either $\mathrm{Pref}(\mathcal{L})$ or $\mathrm{Pref}(\overline{\mathcal{L}})$ are different from Σ^*. Indeed it easily follows from Corollary 9 that this class coincides with $1^{\uparrow}\mathrm{RDEF}$.

In order to complete our characterization, we need to consider the case in which both $\mathcal{L}$ and $\overline{\mathcal{L}}$ are Wheeler and $\mathrm{Pref}(\mathcal{L}) = \mathrm{Pref}(\overline{\mathcal{L}}) = \Sigma^*$. In this case we will generalize the notion of interval on Σ^*. A (classic) *interval* in $(\Sigma^*, <)$ is a set of the form

$$(\alpha, \beta), [\alpha, \beta), (\alpha, \beta], [\alpha, \beta], (\alpha, \infty), [\alpha, \infty),$$

where $\alpha, \beta \in \Sigma^*$ and

$$(\alpha, \beta) = \{\gamma \in \Sigma^* \ : \ \alpha < \gamma < \beta\}, \ \ [\alpha, \beta) = \{\gamma \in \Sigma^* \ : \ \alpha \leq \gamma < \beta\}, \text{ etc...}$$

The words α, β are called the *bound words* of the intervals $(\alpha, \beta), [\alpha, \beta), \ldots$

Following the ideas introduced in [11], we generalize this type of intervals by considering intervals of *finite* words but whose bound words can be periodic in Σ^ω. Since we will compare finite and infinite words co-lexicographically, it is convenient to think of infinite words as "growing" to the left. More precisely, an infinite word σ is depicted as $\sigma = \ldots \sigma_n \sigma_{n-1} \ldots \sigma_1 \sigma_0$, where $\sigma_i \in \Sigma$. Finite words are then seen as infinite words with an infinite occurrence of the character # on the left: $\ldots \# \# \# \sigma_n \ldots \sigma_0$,

where $\# < i$, for all $i \in \Sigma$. The co-lexicographic order $<$ over $\Sigma^* \cup \Sigma^\omega$, extending the co-lexicographic order $<$ over Σ^*, can then be defined by:

$$\sigma < \tau \;\Leftrightarrow \exists n\, (\sigma(n) < \tau(n) \;\wedge\; \forall k < n\, \sigma(k) = \tau(k)),$$

where an infinite sequence of $\#$ is appended to finite words as explained above. Finally, if $\gamma, \delta \in \Sigma^*$ we denote by $\gamma^\omega\delta$ the infinite periodic word[3] $\ldots\gamma\gamma\ldots\gamma\delta$.

We can now consider intervals with bounds that are infinite periodic words. E.g. if $\Sigma = \{1, 2, 3\}$ we have:

$$(2^\omega 3, \infty) \;=\; \{\alpha \in \Sigma^* \;:\; 2^\omega 3 < \alpha\} = \Sigma^* 32^* 3,$$

while $[33, \infty) = \Sigma^* 33$, and $[3321, 31) = \Sigma^* 3321$. As we shall prove later, these generalized intervals are always (regular) Wheeler languages. First we characterize the class DEF using usual (classical) intervals.

Lemma 10. *A regular language $\mathcal{L}$ is in* DEF *if and only if it is a union of a finite number of intervals in $(\Sigma^*, <)$ having finite words or ∞ as bounds.*

Proof. Let $\Sigma = \{1, 2, \ldots, k\}$. We first consider languages of the form $\Sigma^*\beta$. If $\beta = \sigma_1 \ldots \sigma_n$, with $\sigma_j \in \Sigma$, we consider two cases:

1. $\beta \in k^*$; in this case $\Sigma^*\beta = [\beta, \infty)$.
2. $\beta \notin k^*$; in this case let h be the first index from the left such that $\sigma_h = i \neq k$, that is, $\beta = k^{h-1}\; i\; \sigma_{h+1} \ldots \sigma_n$, with $i \neq k$. Define the finite word β' as $\beta' = (i+1)\; \sigma_{h+1} \ldots \sigma_n$. Then $\Sigma^*\beta = [\beta, \beta')$.

Suppose $\mathcal{L} \in$ DEF. Then there are finite sets $F = \{\alpha_1, \ldots \alpha_m\}, G = \{\beta_1, \ldots, \beta_n\}$ such that $\mathcal{L} = F \cup \Sigma^* G$ and, thus, $\mathcal{L} = \{\alpha_1\} \cup \cdots \cup \{\alpha_m\} \cup \Sigma^*\beta_1 \cup \cdots \cup \Sigma^*\beta_n$. Since $\{\alpha_i\} = [\alpha_i]$, for all i, by the previous result $\mathcal{L}$ is a union of a finite number of intervals in $(\Sigma^*, <)$ having finite words or ∞ as bounds.

Conversely, if $\beta \in \Sigma^*$, we first prove that $[\beta, \infty) \in$ DEF. If $\beta = k^j \in k^*$, then $[\beta, \infty) = \Sigma^* k^j \in$ DEF. If $\beta \notin k^*$, then let β' be defined as in (2) above. Then, starting from β, we can reach a word in $k^j \in k^*$ in a finite number of iterations of $(\cdot)'$, obtaining $[\beta, \infty) = [\beta, \beta') \cup [\beta', \beta'') \cup \ldots \cup [k^j, k^\omega) = \Sigma^*\beta \cup \Sigma^*\beta' \cup \cdots \cup \Sigma^* k^j$, proving that $[\beta, \infty) \in$ DEF. Since $[\beta, \infty) \in$ DEF, its complement $[\epsilon, \beta)$ is in DEF as well and the same holds for $[\epsilon, \beta] = [\epsilon, \beta) \cup \{\beta\}$ and $(\beta, \infty) = [\beta, \infty) \setminus \{\beta\}$. Finally, if $\beta < \gamma$ then $(\beta, \gamma) = (\beta, \infty) \cap [\epsilon, \gamma)$ is in DEF, together with all its variations obtained by closing the interval to the left or to the right. Since DEF is closed under finite unions, this proves that a union of a finite number of intervals with finite or ∞ as bounds is in DEF. □

Remark 4. Notice that the usual definition of DEF does not depend on a fixed order of the alphabet. The previous lemma tells us that if $\mathcal{L} \in$ DEF then, whatever the order $<$ is, we can decompose $\mathcal{L}$ as a union of intervals, but these intervals change if we change the order. E.g. if $\Sigma = \{a, b\}$ and $a < b$ then $\Sigma^* a = [a, b)$, while, if $b < a$ then $\Sigma^* a = [a, \infty)$.

In the previous lemma we proved that languages in DEF are those of the form $\bigcup_{i=1}^n I_i$ where I_i are open/closed intervals with bounds $\alpha, \beta \in \Sigma^* \cup \{\infty\}$. We call them *intervals with finite bounds*. We now prove that, by allowing bounds to be also periodic infinite words, we get exactly the class of regular Wheeler languages $\mathcal{L}$ with $\mathrm{Pref}(\mathcal{L}) = \Sigma^*$.

Definition 4.3. The class $\mathrm{DEF}^{\mathrm{lim}}$ is the class of regular languages which are finite unions of intervals in $(\Sigma^+, <)$ with finite or periodic words as bounds[4].

By Lemma 10 we have $\mathrm{DEF} \subseteq \mathrm{DEF}^{\mathrm{lim}}$.

[3] Using the embedding introduced in [11] one such word would map to a periodic rational.

[4] Notice that if $k = max\ \Sigma$ then we can use k^ω instead of ∞ as bound of an interval.

Lemma 11. *The class* $\mathrm{DEF}^{\lim}$ *is closed under boolean operations. Moreover,* $\mathrm{DEF}^{\lim} \subseteq \mathcal{W}(<)$.

Proof. The complement of an interval is either an interval or a union of two intervals. The intersection of a finite number of intervals is an interval. Hence, the complement of a finite union of intervals is a finite unions of intervals. Clearly, the union of two sets that are finite union of intervals is still a finite union of intervals, hence $\mathrm{DEF}^{\lim}$ is closed under boolean operations.

Suppose $\mathcal{L} = \bigcup_{i=1}^{n} I_i$, where the I_i's are intervals. We first prove that $\mathcal{L}$ is regular. We have already proved that intervals with finite bounds are regular. Consider an interval of type $(\gamma^\omega\delta, \infty)$. Then

$$(\gamma^\omega\delta, \infty) = \{\alpha \in \Sigma^* : \gamma^\omega\delta < \alpha\} = \Sigma^* F_1 \cup \Sigma^* F_2 \gamma^* \delta,$$

with $F_1 = \{\, \beta \in \Sigma^* : |\beta| \leq |\gamma\delta|, \gamma\delta < \beta \,\}$, $F_2 = \{\, \beta \in \Sigma^* : |\beta| \leq |\gamma|, \gamma < \beta \,\}$. Hence $(\gamma^\omega\delta, \infty)$ is regular. Since any other interval with bounds which are finite or periodic can be constructed from intervals of this form or with finite bounds using boolean operations, the claim follows.

We are now left to prove that the regular language $\mathcal{L} = \bigcup_{i=1}^{n} I_i$ is Wheeler. If $\mathcal{L}$ is finite, this is obvious. If $\mathcal{L}$ is not finite, then we use Lemma 2. Assume, by way of a contradiction, that there exists a monotone sequence $(\alpha_i)_{i\in\omega}$ which visit alternatively two different states q, q' of the minimum DFA accepting $\mathcal{L}$. Let $\beta \in \Sigma^*$ be such that $\delta(q, \beta) \in F, \delta(q', \beta) \notin F$ (including the case $\delta(q', \beta) = \perp$), and consider the monotone sequence $(\alpha_i\beta)_{i\in\omega}$. This sequence is in $\mathcal{L}$ infinitely often, hence there must be an interval I_k that contains an infinite number of elements of $(\alpha_i\beta)_{i\in\omega}$. Notice that, by definition, an interval contains all words of Σ^* between its bound words. Hence, all elements of the sequence $(\alpha_i\beta)_{i\in\omega}$ will be eventually in I_k and, as a consequence, in $\mathcal{L}$, contradicting the fact that α_i reaches q' for infinitely many i's and $\delta(q', \beta) \notin F$. □

Remark 5. In Lemma 11 we have just proved that any finite union of intervals in Σ^*, with finite or periodic words as bounds, is a Wheeler language. It is not possible to generalize this result to intervals in $\mathrm{Pref}(\mathcal{L})$, e.g. the language $\mathcal{L} = 13^*1 \cup 23^*2 \cup \{2\}$ is the union of two intervals in $\mathrm{Pref}(\mathcal{L})$, that is,

$$\mathcal{L} = \{\, \alpha \in \mathrm{Pref}(\mathcal{L}) : 11 \leq \alpha < 3^\omega 1 \,\} \cup \{\, \alpha \in \mathrm{Pref}(\mathcal{L}) : 2 \leq \alpha < 3^\omega 2 \,\}$$

(notice that e.g. $11 < 21 < 3^\omega 1$ but $21 \notin \mathrm{Pref}(\mathcal{L})$), but $\mathcal{L}$ is not Wheeler.

Remark 6. There is an important difference, when Wheelerness is concerned, between intervals with finite or ∞ bounds and intervals with periodic bounds. If a language $\mathcal{L}$ is a finite union of classical intervals and we change the order, then $\mathcal{L}$ is still a finite union of classical intervals w.r.t. the new order (because it is in DEF, see Remark 4). In other words, the class DEF is not sensitive to the underlying order of the alphabet: although our characterization seems to depend from the order, the usual definition of the class does not. This independence does not hold for intervals with periodic bounds. E.g. consider the alphabet $\Sigma = \{1, 2, 3\}$ and the language $\mathcal{L} = \Sigma^*12^*$, which is in $\mathrm{DEF}^{\lim}$ if the order is the standard $1 < 2 < 3$ because $\Sigma^*12^* = [1, 2^\omega)$. In this order there are two infinite sequences of words converging, from below and above, respectively, to 2^ω—namely, $\alpha 12^n$ and $\alpha 32^n$, for $\alpha \in \Sigma^*$ and $n \in \mathbb{N}^+$.

The two sequences, however, must reach different states (one accepting and the other not accepting) in the automaton recognizing $\mathcal{L}$, and cannot be merged into a unique monotone sequence: hence, there is no violation of Lemma 2 . If, instead, we consider the order $1 < 3 < 2$, the language is not Wheeler because with this order the two sequence above can be merged into a unique monotone sequence

$$\alpha 12 < \alpha 32 < \alpha 122 < \alpha 322 < \cdots < \alpha 12^n < \alpha 32^n < \alpha 12^{n+1} < \alpha 32^{n+1} \ldots$$

violating Lemma 2. Notice that, if the order is $1 < 3 < 2$, then Σ^*12^* is not an interval in Σ^*, because $1, 12 \in \Sigma^*12^*$ while $13 \notin \Sigma^*12^*$, although $1 < 13 < 12$ holds.

Lemma 12 ([11])**.** *Suppose* $\mathrm{Pref}(\mathcal{L}) = \Sigma^*$. *Then* $\mathcal{L} \in \mathcal{W}(<) \;\Rightarrow\; \mathcal{L} \in \mathrm{DEF}^{\lim}$.

Proof. This result was already proved in [11]. In this paper it is proved that, given a Wheeler DFA, the set of words ending in any given state constitute an interval of $\mathrm{Pref}(\mathcal{L})$ bounded by finite or periodic words. If $\mathcal{L}$ is Wheeler and $\mathrm{Pref}(\mathcal{L}) = \Sigma^*$, then there exists a WDFA D recognizing $\mathcal{L}$ and $\mathcal{L}$ is a union of the Σ^*-interval corresponding to final states. □

Using Lemma 7 and Lemma 12 we finally obtain the following.

Theorem 13. *Let $\mathcal{L}$ be a regular language. Then*

$$\mathcal{L}, \overline{\mathcal{L}} \in W(<) \;\Leftrightarrow\; \mathcal{L} \in 1^{\uparrow}\mathrm{RDEF} \cup \mathrm{DEF}^{\lim}$$

Proof. ($\Rightarrow$) If either $\mathrm{Pref}(\mathcal{L})$ or $\mathrm{Pref}(\overline{\mathcal{L}})$ are different from Σ^* then $\mathcal{L}, \overline{\mathcal{L}}$ are both in $1^{\uparrow}\mathrm{RDEF}$ by Lemma 7 and Lemma 5. If $\mathrm{Pref}(\mathcal{L}) = \mathrm{Pref}(\overline{\mathcal{L}}) = \Sigma^*$, then $\mathcal{L}, \overline{\mathcal{L}}$ are both in $\mathrm{DEF}^{\lim}$ by Lemma 12 and Lemma 11.

($\Leftarrow$) If $\mathcal{L} \in 1^{\uparrow}\mathrm{RDEF}$ then $\overline{\mathcal{L}} \in 1^{\uparrow}\mathrm{RDEF}$ and both $\mathcal{L}$ and $\overline{\mathcal{L}}$ are Wheeler according to Lemma 6. If $\mathcal{L} \in \mathrm{DEF}^{\lim}$ then $\overline{\mathcal{L}} \in \mathrm{DEF}^{\lim}$ and both $\mathcal{L}$ and $\overline{\mathcal{L}}$ are Wheeler by Lemma 11. □

Notice that $1^{\uparrow}\mathrm{RDEF} \cap \mathrm{DEF}^{\lim}$ is not limited to finite and cofinite languages. Indeed, the language $\overline{\mathcal{L}}$ of Example 4.1 belongs to the intersection $1^{\uparrow}\mathrm{RDEF} \cap \mathrm{DEF}^{\lim}$. In fact, it belongs to $1^{\uparrow}\mathrm{RDEF}$ because $\overline{\mathcal{L}} = 1^{\uparrow}(\{2, 12\}\{1, 2\}^* + \epsilon)$ and belongs to $\mathrm{DEF}^{\lim}$ by Lemma 12, since it is Wheeler and $\mathrm{Pref}(\overline{\mathcal{L}}) = \Sigma^*$. Moreover, notice that Lemma 7 and Lemma 12 fail to give a complete characterization of Wheeler language only for the language $\mathcal{L}$ such that $\mathcal{L} \in \mathrm{W}(<)$, $\overline{\mathcal{L}} \notin \mathrm{W}(<)$, and $\mathrm{Pref}(\mathcal{L}) \neq \Sigma^*, \mathrm{Pref}(\overline{\mathcal{L}}) = \Sigma^*$.

5. Conclusions

In this paper we proved that the class of Wheeler languages $\mathcal{L}$ whose complement $\overline{\mathcal{L}}$ is also Wheeler can be fully characterized and shows interesting features. On the one hand, such class is best described starting from the classic classes DEF and RDEF: any $\mathcal{L} \in \mathrm{DEF} \cup \mathrm{RDEF}$ is certainly Wheeler with complement also Wheeler. On the other hand, however, both DEF and RDEF need a "closer look" for a complete characterization: both must be extended to capture (exactly) Wheeler languages with Wheeler complement. For any $\mathcal{L} \in \mathrm{RDEF}$ also $1^{\uparrow}\mathcal{L}$ is Wheeler with Wheeler complement. Moreover, and probably more interestingly, the class DEF must also be extended to a class $\mathrm{DEF}^{\lim}$ that can be seen as a partition of Σ^* into a collection of open/closed intervals. Both cases need further investigation, in particular in view of the possibility of relativizing this analysis to the (much more general) case of *partial* orderings of states.

Acknowledgments

The authors thank Ruben Becker and Nicola Prezza for helpful discussions and comments.

Funding. Giuseppa Castiglione is Supported by Project "ACoMPA" (CUP B73C24001050001) funded by the NextGeneration EU programme PNRR MUR M4 C2 Inv. 1.5 – Project ECS00000017 Tuscany Health Ecosystem (Spoke 6), CUP Master B63C22000680007.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] T. Gagie, G. Manzini, J. Sirén, Wheeler graphs: a framework for BWT-based data structures, Theoretical Computer Science 698 (2017) 67 – 78. doi:10.1016/j.tcs.2017.06.016, algorithms, Strings and Theoretical Approaches in the Big Data Era (In Honor of the 60th Birthday of Professor Raffaele Giancarlo).

[2] J. Alanko, N. Cotumaccio, N. Prezza, Linear-time minimization of wheeler dfas, in: 2022 Data Compression Conference (DCC), 2022, pp. 53–62. doi:10.1109/DCC52660.2022.00013.

[3] J. Alanko, G. D'Agostino, A. Policriti, N. Prezza, Wheeler languages, Inf. Comput. 281 (2021) 104820. URL: https://doi.org/10.1016/j.ic.2021.104820. doi:10.1016/J.IC.2021.104820.

[4] R. Becker, D. Cenzato, S.-H. Kim, B. Kodric, A. Policriti, N. Prezza, Optimal wheeler language recognition, in: International Symposium on String Processing and Information Retrieval, Springer, 2023, pp. 62–74.

[5] N. Cotumaccio, G. D'Agostino, A. Policriti, N. Prezza, Co-lexicographically ordering automata and regular languages - part I, J. ACM 70 (2023) 27:1–27:73. doi:10.1145/3607471.

[6] G. Castiglione, A. Restivo, Completing wheeler automata, in: U. de'Liguoro, M. Palazzo, L. Roversi (Eds.), Proceedings of the 25th Italian Conference on Theoretical Computer Science, Torino, Italy, September 11-13, 2024, volume 3811 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 120–132. URL: https://ceur-ws.org/Vol-3811/paper060.pdf.

[7] R. Becker, G. Castiglione, G. D'Agostino, A. Policriti, N. Prezza, A. Restivo, B. Riccardi, Universally wheeler languages, CoRR abs/2504.19537 (2025). URL: https://doi.org/10.48550/arXiv.2504.19537. doi:10.48550/ARXIV.2504.19537. arXiv:2504.19537.

[8] S. C. Kleene, Representation of Events in Nerve Nets and Finite Automata, Princeton University Press, Princeton, 1956, pp. 3–42. doi:doi:10.1515/9781400882618-002.

[9] J. A. Brzozowski, B. Li, D. Liu, Syntactic complexities of six classes of star-free languages, J. Autom. Lang. Comb. 17 (2012) 83–105. doi:10.25596/JALC-2012-083.

[10] J. Sakarovitch, Elements of automata theory, Cambridge university press, 2009.

[11] G. Manzini, A. Policriti, N. Prezza, B. Riccardi, The rational construction of a wheeler DFA, in: 35th Annual Symposium on Combinatorial Pattern Matching, CPM 2024, June 25-27, 2024, Fukuoka, Japan, volume 296 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024, pp. 23:1–23:15. doi:10.4230/LIPICS.CPM.2024.23.