A Kleene Theorem for Languages of Words Indexed by Linear Orderings

Alexis Bès, Olivier Carton

A Kleene Theorem for Languages of Words Indexed by Linear Orderings

Alexis Bès
LACL, Université Paris XII-Val de Marne,
Email: bes@univ-paris12.fr,
Url: http://www.univ-paris12.fr/lacl/bes/

Olivier Carton
LIAFA, Université Paris 7
Email: Olivier.Carton@liafa.jussieu.fr,
Url: http://www.liafa.jussieu.fr/~carton/

March 20, 2006

Abstract

In a preceding paper, Bruyère and Carton introduced automata, as well as rational expressions, which allow to deal with words indexed by linear orderings. A Kleene-like theorem was proved for words indexed by countable scattered linear orderings. In this paper we extend this result to languages of words indexed by all linear orderings.

1 Introduction

One of the fundamental results in automata theory is Kleene's theorem [16] which asserts the equivalence between sets of words accepted by automata and set of words described by rational expressions. During the past fifty years Kleene's theorem has been extended to various notions of infinite words, as well as structures like trees, pictures, and traces.

In [5, 3], Bruyère and Carton introduce automata and rational expressions for words on linear orderings. These notions unify naturally previously defined notions for finite words, left- and right-infinite words, bi-infinite words, and ordinal words. They also prove that a Kleene-like theorem holds when the orderings are restricted to countable scattered linear orderings; recall that a linear ordering is scattered if it does not contain any dense sub-ordering. This result extends Kleene's theorem for finite words [16], infinite words [6, 17], bi-infinite words [13, 18] and ordinal words [8, 10, 27]. Since [3], the study of automata on linear orderings was carried on in several papers, that address the emptiness problem and the containment problem for languages [26, 9], as well as the classification of rational languages with respect to the rational operations needed to describe them [4]. More recently, Carton and Rispal [21] proved that regular languages of words over countable scattered linear orderings are closed under complementation.

In this paper we come back to Kleene's theorem, and show that the assumption that the linear orderings are countable and scattered, is not necessary. When all linear orderings are considered instead of countable and scattered ones, no change has to be made to the notion of automata already introduced in [3]. However the set of operators for the rational expressions has to be extended in order to deal with words indexed by a dense linear ordering. To cope with this issue, we add a *shuffle* operator for languages, which is a variant of the classical shuffle operation on linear orderings. A similar (but not equivalent) notion of shuffle for languages was already considered in [11, 15, 24]. This operator allows to extend the definition of rational languages of words, and to prove a general Kleene-like theorem.

Words indexed by a countable linear ordering were first considered in [11], where they were introduced as frontiers of labeled binary trees. Some kind of rational expressions were studied in [11, 15, 24], which lead to a characterization of words which are frontiers of regular trees.

Other related works can be found in the area of specification and verification of real-time systems. Indeed, words indexed by $\mathbb{R}$ (or other linear orderings) appear as a simple and natural way to model the behavior of a finite state real-time system. For example, ordinal words (called Zeno words) were recently considered as modeling infinite sequences of actions which occur in a finite interval of time [14, 2]. While the intervals of time are finite, infinite sequences of actions can be concatenated. A Kleene's theorem already exists for standard timed automata (where infinite sequences of actions are supposed to generate divergent sequences of times) [1]. In [2], automata considered by Choueka and Wojciechowski are adapted to Zeno words. A kind of Kleene's theorem is proved, that is, the class of Zeno languages is the closure under an operation called refinement of the class of languages accepted by standard timed automata. More recently, Dima introduces a notion of real-time automata [12] that captures a class of timed languages which is closed under complementation and for which a Kleene's theorem is proved. Let us finally mention the paper [20] which introduces star-free expressions for words indexed by $\mathbb{R}$, and shows that star-free languages of words indexed by $\mathbb{R}$ coincide with languages definable in some first-order logic, extending McNaughton-Papert theorem.

The paper is organized as follows: we recall in Sect. 2 some useful definitions related to linear orderings. Sections 3 and 4 respectively introduce rational expressions and automata for words over linear orderings. Section 5 states the main theorem and provides a few examples. Sections 6 and 7.

2 Linear Orderings

In this section we recall useful definitions and results about linear orderings. A good reference on the subject is Rosenstein's book [22].

A *linear ordering* J is an ordering $<$ which is total, that is, for any $j \neq k$ in J, either $j < k$ or $k < j$ holds. Given a linear ordering J, we denote by $-J$ the *backwards* linear ordering obtained by reversing the ordering relation. For instance, $-\omega$ is the backwards linear ordering of ω which is used to index the so-called left-infinite words.

The sum of orderings is concatenation. let J and K_j for $j \in J$, be linear orderings. The linear ordering $\sum_{j \in J} K_j$ is obtained by juxtaposition of the orderings K_j with respect to J. More formally, the *sum* $\sum_{j \in J} K_j$ is the set L of all pairs (k, j) such that $k \in K_j$. The relation $(k_1, j_1) < (k_2, j_2)$ holds iff $j_1 < j_2$ or ($j_1 = j_2$ and $k_1 < k_2$ in K_{j_1}). The sum of two orderings K_1 and K_2 is denoted $K_1 + K_2$.

Two elements j and k of a linear ordering J are called *consecutive* if $j < k$ and if there is no element $i \in J$ such that $j < i < k$. An ordering is *dense* if it contains no pair of consecutive elements. More generally, a subset $K \subset J$ is *dense* in J if for any $j, j' \in J$ such that $j < j'$, there is $k \in K$ such that $j < k < j'$.

The notion of a cut is needed to define a path in an automaton. A *cut* of a linear ordering J is a pair (K, L) of intervals such that $J = K \cup L$ and such that for any $k \in K$ and $l \in L$, $k < l$. The set of all cuts of the ordering J is denoted by $\hat{J}$. This set $\hat{J}$ can be linearly ordered by the relation defined by $c_1 < c_2$ iff $K_1 \subsetneq K_2$ for any cuts $c_1 = (K_1, L_1)$ and $c_2 = (K_2, L_2)$. This linear ordering can be extended to $J \cup \hat{J}$ by setting $j < c_1$ whenever $j \in K_1$ for any $j \in J$.

The consecutive elements of $\hat{J}$ deserve some attention. For any element j of J, define two cuts c_j^- and c_j^+ by $c_j^- = (K, \{j\} \cup L)$ and $c_j^+ = (K \cup \{j\}, L)$ where $K = \{k \mid k < j\}$ and $L = \{k \mid j < k\}$. It can be easily checked that the pairs of consecutive elements of $\hat{J}$ are the pairs of the form (c_j^-, c_j^+).

An ordering J is *complete* if for any cut (K, L) such that $K \neq \varnothing$ and $L \neq \varnothing$, either K has a greatest element or L has a least element.

3 Words and rational expressions

Given a finite alphabet A, a *word* $(a_j)_{j \in J}$ is a function from J to A which maps any element j of J to a letter a_j of A. We say that J is the *length* $|x|$ of the word x. For instance, the *empty word* ε is indexed by the empty linear ordering $J = \varnothing$. Usual finite words are the words indexed by finite orderings $J = \{1, 2, \ldots, n\}$, $n \geq 0$. A word of length $J = \omega$ is usually called an ω-word or an infinite word. A word of length $\zeta = -\omega + \omega$ is a sequence $\ldots a_{-2} a_{-1} a_0 a_1 a_2 \ldots$ of letters which is usually called a bi-infinite word.

The sum operation on linear orderings leads to a notion of product of words as follows. Let J and K_j for $j \in J$, be linear orderings. Let $x_j = (a_{k,j})_{k \in K_j}$ be a word of length K_j, for any $j \in J$. The *product* $\prod_{j \in J} x_j$ is the word z of length $L = \sum_{j \in J} K_j$ equal to $(a_{k,j})_{(k,j) \in L}$. For instance, the word $a^\zeta = b^{-\omega} a^\omega$ of length ζ is the product of the two words $b^{-\omega}$ and a^ω of length $-\omega$ and ω respectively.

We now recall the notion of rational set of words on linear orderings as defined in [3]. The rational operations include of course the usual Kleene operations for finite words which are the union $+$, the concatenation $\cdot$ and the star operation $*$. They also include the omega iteration ω usually used to construct ω-words and the ordinal iteration $\sharp$ introduced by Wojciechowski [27] for ordinal words. Four new operations are also needed: the backwards omega iteration $-\omega$, the backwards ordinal iteration $-\sharp$, a binary operation denoted $\diamond$ which is a kind of iteration for all orderings, and finally a shuffle operation which allows to deal with dense linear orderings.

Let us recall the already known rational operations. We respectively denote by $\mathcal{N}$, $\mathcal{O}$ and $\mathcal{L}$ the classes of finite orderings, the class of all ordinals and the class of all linear orderings. For an ordering J, we denote by $\hat{J}^*$ the set $\hat{J} \setminus \{(\varnothing, J), (J, \varnothing)\}$ where $(\varnothing, J)$ and $(J, \varnothing)\}$ are the first and last cut. Given two sets X and Y of words, define

$$\begin{array}{rcl} X + Y & = & \{z \mid z \in X \cup Y\}, \\ X \cdot Y & = & \{x \cdot y \mid x \in X, y \in Y\}, \\ X^* & = & \{\prod_{j \in \{1,\ldots,n\}} x_j \mid n \in \mathcal{N}, x_j \in X\}, \\ X^\omega & = & \{\prod_{j \in \omega} x_j \mid x_j \in X\}, \\ X^{-\omega} & = & \{\prod_{j \in -\omega} x_j \mid x_j \in X\}, \\ X^\sharp & = & \{\prod_{j \in \alpha} x_j \mid \alpha \in \mathcal{O}, x_j \in X\}, \\ X^{-\sharp} & = & \{\prod_{j \in -\alpha} x_j \mid \alpha \in \mathcal{O}, x_j \in X\}, \\ X \diamond Y & = & \{\prod_{j \in J \cup \hat{J}^*} z_j \mid J \in \mathcal{L}, z_j \in X \text{ if } j \in J \text{ and } z_j \in Y \text{ if } j \in \hat{J}^*\}. \end{array}$$

We use the notation $X^\diamond$ as an abbreviation for $(X \diamond \varepsilon) + \varepsilon$.

We now define the new shuffle operation which is needed to deal with dense orderings.

Definition 1 *Let A be a finite alphabet, $n \geq 1$, and $L_1, \ldots, L_n \subseteq A^\diamond$. We define the shuffle of $L_1, \ldots, L_n$, and denote by $\mathrm{sh}(L_1, \ldots, L_n)$ the set of words $w \in A^\diamond$ that can be written as $w = \prod_{j \in J} w_j$ where*

- *J is a complete linear ordering without first and last element;*
- *there exists a partition $(J_1, \ldots, J_n)$ of J such that all J_i's are dense in J, and for every $j \in J$, if $j \in J_k$ then $w_j \in L_k$.*

Let us remark that our definition of shuffle slightly differs from others, e.g. from [15, 24], because in Definition 1 we assume that J is a *complete* dense ordering.

Only countable orderings are considered in [15, 24]. Recall that $\mathbb{Q}$ is the unique countable and dense ordering without first and last element. Their definition of the shuffle operation is based on a partition $(J_1, \ldots, J_n)$ of $\mathbb{Q}$ into dense subsets $J_1, \ldots, J_n$. Then points of each J_i are substituted by words from L_i as we do. Our definition is not a straightforward generalization of this shuffle because $\mathbb{Q}$ is of course not complete. Actually the assumption that J is complete yields a more general notion of shuffle. The completion of $\mathbb{Q}$ yields the ordering $\mathbb{R}$. If each point of $\mathbb{R} \setminus \mathbb{Q}$ is substituted by the empty word, one obtains a shuffle in the sense of [15, 24]. This shows that our rational expression $\mathrm{sh}(L_1, \ldots, L_n, \varepsilon)$ corresponds to the shuffle of languages $L_1, L_2, \ldots, L_n$ in the sense of [15, 24].

An abstract *rational expression* is a well-formed term of the free algebra over $\{\varnothing\} \cup A$ with the symbols denoting the rational operations as function symbols. Each rational expression denotes a set of words which is inductively defined by the above definitions of the rational operations. A set of words is *rational* if it can be denoted by a rational expression. As usual, the dot denoting concatenation is omitted in rational expressions.

Example 2 Consider the word $w = (w_r)_{r \in \mathbb{R}}$ of length $\mathbb{R}$ over the alphabet $A = \{a, b\}$, defined by $w_r = a$ if $r \in \mathbb{Q}$, and $w_r = b$ otherwise. Then it is not difficult to check that $w \in \mathrm{sh}(a, b)$. Consider now the word $w' = (w'_q)_{q \in \mathbb{Q}}$ of length $\mathbb{Q}$ over the alphabet A, defined by $w'_q = a$ if $q \in \{m/2^n \mid m \in \mathbb{Z}, n \in \mathbb{N}\}$, and $w'_q = b$ otherwise. It can be checked that $w' \in \mathrm{sh}(a, b, \varepsilon)$ but that $w' \notin \mathrm{sh}(a, b)$ because $\mathbb{Q}$ is not complete.

Example 3 The rational expression $a^*(\varepsilon + \mathrm{sh}(a^*, \varepsilon))a^*$ denotes the set of words (over the unary alphabet $\{a\}$) whose length is an ordering containing no infinite sequence of consecutive elements. It is clear that the length of any word denoted by this expression cannot contain an infinite sequence of consecutive elements. Conversely, let J be such an ordering. Define the equivalence relation $\sim$ on J by $x \sim y$ iff there are finitely many elements between x and y. The classes of $\sim$ are then finite intervals. Furthermore the ordering of these intervals must be a dense ordering with possibly a first and a last element. This completes the converse.

Example 4 The rational expression $(\varepsilon + \mathrm{sh}(a)) \diamond a$ denotes the set of words (over the unary alphabet $\{a\}$) whose length is a complete ordering. The shuffle operator is defined using complete orderings and the ordering $\hat{J}$ is always complete. It follows from these two facts that the length of any word denoted by this expression is complete. Conversely, let J be a complete orderings. Define the equivalence relation $\sim$ on J by $x \sim y$ iff there is an open dense interval containing both x and y. Each class of $\sim$ is either a singleton or an open dense interval. Let K be the ordering of the singleton classes and let L_0 be the ordering of the dense classes. Let L_1 be the ordering of pairs of consecutive elements in K and let L be $L_0 \cup L_1$ equipped with the natural ordering. It can be shown that $K = \hat{L}$. This gives the expression $(\varepsilon + \mathrm{sh}(a)) \diamond a$ where ε is due to L_1, $\mathrm{sh}(a)$ to L_0 and a to K. This completes the converse.

4 Automata

In this section, we recall the definition given in [3] for automata accepting words on linear orderings. As already noted in [3], this definition is actually suitable for all linear orderings.

Automata accepting words on linear orderings are classical finite automata equipped with limit transitions. They are defined as $\mathcal{A} = (Q, A, E, I, F)$, where Q denotes the finite set of states, A is a finite alphabet, and I, F denote the set of initial and final states, respectively. The set E consists in three types of transitions: the usual *successor* transitions in $Q \times A \times Q$, the *left limit* transitions which belong to $2^Q \times Q$ and the *right limit* transitions which belong to $Q \times 2^Q$. A left (respectively right) limit transition $(P, q) \in 2^Q \times Q$ (respectively, $(q, P) \in Q \times 2^Q$) will usually be denoted by $P \to q$ (respectively $q \to P$).

We say that a transition *leaves* a state q if it either a successor transition $q \xrightarrow{a} p$ for some state p or a right limit transition $q \to P$ for some subset P of states. We say that it *enters* a state q if it is either a successor transition $p \xrightarrow{a} q$ or a left limit transition $P \to q$. The sets of transitions leaving and entering a state q are respectively denoted by $\mathrm{Out}(q)$ and $\mathrm{In}(q)$.

We sometimes write that an automaton $\mathcal{A}$ has transitions $P_1, \ldots, P_m \to q_1, \ldots, q_n$ when $\mathcal{A}$ has all left limit transitions $P_i \to q_j$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. Analogously we shall use the notation $q_1, \ldots, q_n \to P_1, \ldots, P_m$ for right limit transitions.

A word $x = (a_j)_{j \in J}$ of length J is accepted by $\mathcal{A}$ if it is the label of a successful path. A *path* γ is a sequence of states $\gamma = (q_c)_{c \in \hat{J}}$ of length $\hat{J}$ verifying the following conditions. For two consecutive states in γ, there must be a successor transition labeled by the letter in between. For a state $q \in \gamma$ which has no predecessor on γ, there must be a left limit transition $P \to q$ where P is the limit set of γ on the left of q. Right limit transitions are used similarly when q has no successor on γ.

Since a sequence of states indexed by $\hat{J}$ is actually a function from $\hat{J}$ into Q, we sometimes use a functional notation and the state q_c of a path γ is also denoted by $\gamma(c)$.

Observe that the ordering $\hat{J}$ always has a first element and a last element, namely the cuts $c_{\min} = (\varnothing, J)$ and $c_{\max} = (J, \varnothing)$. For any cut $c \in \hat{J}$, define the sets $\lim_{c^-} \gamma$ and $\lim_{c^+} \gamma$ as follows:

$$\lim_{c^-} \gamma = \{q \in Q \mid \forall c' < c \;\; \exists k \quad c' < k < c \text{ and } q = q_k\},$$
$$\lim_{c^+} \gamma = \{q \in Q \mid \forall c < c' \;\; \exists k \quad c < k < c' \text{ and } q = q_k\}.$$

A sequence $\gamma = (q_c)_{c \in \hat{J}}$ of states is an accepting path for the word $x = (a_j)_{j \in J}$ if the following conditions are fulfilled. For any pair (c_j^-, c_j^+) of consecutive cuts of J, the automaton must have the successor transition $q_{c_j^-} \xrightarrow{a_j} q_{c_j^+}$. For any cut $c \neq c_{\min}$ which has no predecessor in $\hat{J}$, $\lim_{c^-} \gamma \to q_c$ must be a left limit transition. For any cut $c \neq c_{\max}$ in $\hat{J}$ which has no successor, $q_c \to \lim_{c^+} \gamma$

must be a right limit transition. A path is *successful* if its first state $q_{c_{\min}}$ is initial and its last state $q_{c_{\max}}$ is final.

Example 5 Let $A = \{a, b\}$. The automata $\mathcal{A}$ pictured in Fig. 1 has two successor transitions, three left limit transitions and three right limit transitions. State 0 is the only initial state, and state 5 is the only final state.

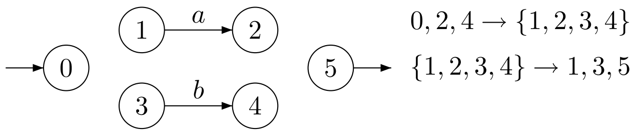


Figure 1: Automaton accepting $\mathrm{sh}(a, b)$

Let us show that this automata accepts words in $\mathrm{sh}(a, b)$. Consider indeed a word $w = (w_j)_{j\in J}$, and assume first that w is accepted by $\mathcal{A}$. Let $\gamma = (q_c)_{c\in \hat{J}}$ be a successful path labeled by w. The ordering J must be dense since there are no consecutive transitions in $\mathcal{A}$. It must also be complete since there is no state with incoming left limit transitions and leaving right limit transitions. Occurrences of both a and b must be dense in J since all limit transitions involve the four states $\{1, 2, 3, 4\}$. Finally, J cannot have a first or a last element. Indeed, the only transition leaving state 0 is a limit one, and similarly for the only transition entering state 5.

Conversely, let $w = (w_j)_{j\in J}$ be a word indexed by a complete ordering J such that occurrences of both a and b are dense in J. Since J is complete, any cut of J (apart from $c_{\min}$ and $c_{\max}$) are either preceded or followed by a letter. Then the sequence $\gamma = (q_c)_{c\in \hat{J}}$ defined as follows is a successful path labeled by w.

- $q_{c_{\min}} = 0$,
- $q_{c_{\max}} = 5$,
- $q_c = 1$ if c is followed by an a and $q_c = 2$ if c is preceded by an a,
- $q_c = 3$ if c is followed by a b and $q_c = 4$ if c is preceded by a b.

The construction of an automaton accepting $\mathrm{sh}(L_1, \ldots, L_n)$ from automata accepting $L_1, \ldots, L_n$ is a straightforward generalization of the automaton for $\mathrm{sh}(a, b)$ which is pictured in Fig. 1.

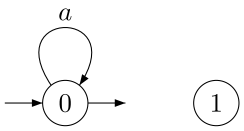
5 Rational expressions vs automata

In this section we state the main theorem of the paper, i.e. that automata and rational expressions define the same languages of words over linear orderings.

Theorem 6 *A set of words over linear orderings is rational iff it is recognizable.*

This result was proved in [3] for the restricted case of countable scattered linear orderings. Many arguments still hold in the general case. As in [3], the *only if* part of the proof relies upon an induction on the rational expression. The only modification with respect to the proof of [3] is that we have to show that the shuffle of rational languages is rational. For the *if* part of the proof, we use as in [3] Yamada's classical technique, i.e. an induction on the set of states visited by a successful path of the automaton. A key ingredient is the use of successive condensations of linear orderings.

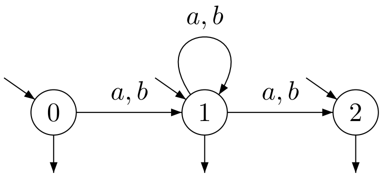
We illustrate the theorem with a few examples, over the alphabet $A = \{a, b\}$.



$$0, 1 \to \{0, 1\}$$
$$\{0, 1\} \to 0, 1$$

Figure 2: Automaton accepting $a^*(\varepsilon + \mathrm{sh}(a^*, \varepsilon))a^*$

Example 7 The automaton pictured in Fig. 2 accepts the set $a^*(\varepsilon + \mathrm{sh}(a^*, \varepsilon))a^*$ of words already considered in Example 3.

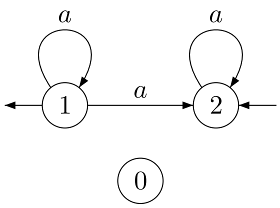


$$2 \to \{1\}, \{0, 1\}, \{1, 2\}, \{0, 1, 2\}$$
$$\{1\}, \{0, 1\}, \{1, 2\}, \{0, 1, 2\} \to 0$$

Figure 3: Automaton accepting words with a complete length

Example 8 The automaton pictured in Fig. 3 accepts words over $\{a, b\}$ whose length is a complete ordering. Since all limit transitions enter state 0 and leave state 2, the length of an accepted word must be complete. Conversely, let x be a word of length J where J is a complete ordering. Define the path γ which maps any cut (K, L) of $\hat{J}$ to 0 if K has no greatest element, to 2 if L has no least element and to 1 otherwise. It is pure routine to check that this defines an accepting path for x.

Example 9 The automaton pictured in Fig. 4 accepts words over $\{a\}$ whose length is a non scattered ordering. Let $w = (w_j)_{j \in J}$ be a word labeling a successful path γ in $\mathcal{A}$. Let K be the set of positions j such that w_j is read by the transition $1 \xrightarrow{a} 2$ in γ. It can be checked that K is a dense subordering of J. Therefore, the ordering J is not scattered.



$1 \to \{1\}$

$\{1\} \to 1$

$2 \to \{2\}$

$\{2\} \to 2$

$0, 2 \to \{1, 2\}, \{0, 1, 2\}$

$\{1, 2\}, \{0, 1, 2\} \to 0, 1$

Figure 4: Automaton accepting words with a non scattered length

For the converse, recall that each linear ordering J can be written $J = \sum_{k \in K} J_k$ where each ordering J_k is scattered and the ordering K is either the one-element ordering if J is scattered or a dense ordering [22, chap. 4]. From this decomposition, a word w whose length is not scattered is equal to a product $\prod_{k \in K} w_k$ where K is a dense ordering. Then a path γ_k from state 1 to state 2 labeled by w_k can be constructed as follows. Let J_k be the length of w_k and let z_k be an arbitrarily chosen element of J_k. Any cut of J_k before z_k is mapped to state 1 and any cut after z_k is mapped to state 2. A path labeled by w is finally constructed as follows. Any cut inside some w_k is mapped to the corresponding state in γ_k and any remaining gap is mapped to state 0.

6 From rational expressions to automata

In this section we prove that every rational set of words is accepted by some automaton. The proof goes by induction on the rational expression denoting the set. For each rational operation we must describe a corresponding construction for the automata. The constructions given in [5] for the operators $\cup$, $\cdot$, *, $^\omega$, $^{-\omega}$, $^\sharp$, $^{-\sharp}$ and $\diamond$, do not depend on any particular assumption on the linear orderings, and thus remain correct in our context. Therefore we only have to provide a construction for the shuffle operation.

We shall work with *normalized automata*, i.e automata which have a unique initial state i and a unique final state $f \neq i$, and have no transition which enters i or leaves f. Note that these conditions imply that the states i and f can only occur as the first state and the last state of a path. Therefore the transitions of the form $P \to q$ or $q \to P$ where P contains i or f cannot occur in a path. In the sequel, we assume that a normalized automaton does not have transitions of the form $P \to q$ or $q \to P$ where P contains i or f.

The following lemma, the proof of which can be found in [3], states that the empty word can be added or removed without changing recognizability. Furthermore a recognizable set which does not contain the empty word can be accepted by a normalized automaton. Note that this condition is necessary since a normalized automaton cannot accept the empty word.

LEMMA 10 *Let X be a set of words. The set X is recognizable iff $X + \varepsilon$ is recognizable. Furthermore if $\varepsilon \notin X$, then X can be recognizable by a normalized automaton.*

We have to show that for every $n \geq 1$, if $X_1, \ldots, X_n$ are recognizable then $\mathrm{sh}(X_1, \ldots, X_n)$ is also recognizable.

We first assume that the empty word does not belong to any of the X_i's. Then by our assumption the languages $X_1, \ldots, X_n$ are accepted by the normalized automata $\mathcal{A}_1, \ldots, \mathcal{A}_n$, respectively. Suppose that $\mathcal{A}_t = (Q_t, E_t, \{i_t\}, \{f_t\})$ for every $t \in \{1, \ldots, n\}$, and $Q_i \cap Q_j = \varnothing$ whenever $i \neq j$. Consider the automaton $\mathcal{A}$ obtained by juxtaposition of $\mathcal{A}_1, \ldots, \mathcal{A}_n$, by adding two new states i, f which are the initial state and the final state of $\mathcal{A}$ (respectively), and by adding all transitions of the form $P \to i_1, \ldots, i_n, f$ and $f_1, \ldots, f_n, i \to P$, where $1 \leq k \leq n$ and $P \supseteq \{i_1, \ldots, i_n, f_1, \ldots, f_n\}$. The construction is pictured in Figure 5.

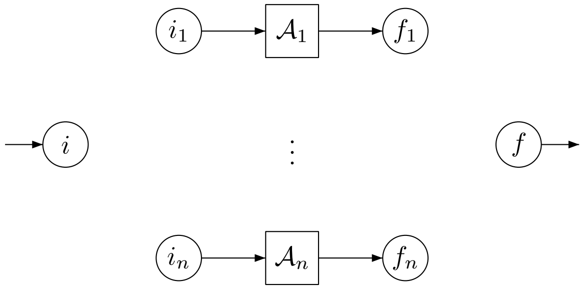


Figure 5: Automaton for $\mathrm{sh}(X_1, \ldots, X_n)$

Let us show that $\mathcal{A}$ accepts exactly $\mathrm{sh}(X_1, \ldots, X_n)$. First assume that $w \in \mathrm{sh}(X_1, \ldots, X_n)$, i.e. $w = \prod_{j \in J} w_j$ where J is a complete linear ordering without first and last element, and $(J_1, \ldots, J_n)$ is a partition of J such that all J_i's are dense in J, and for every $j \in J$, if $j \in J_k$ then $w_j \in X_k$. Then every w_j is the label of a successful path γ_j in some $\mathcal{A}_t$ (t is such that $j \in J_t$). It is clear that the paths γ_j can be concatenated, with an additional state i at the beginning and an additional state f at the end, to form a successful path in $\mathcal{A}$.

Conversely if $w = (x_k)_{k \in K}$ is a word indexed by the linear ordering K and is the label of a successful path $\gamma = (\gamma_k)_{k \in \hat{K}}$ in $\mathcal{A}$, then w can be written as $w = \prod_{j \in J} w_j$ where each w_j is the label of a subpath of γ of the form $(i_t, \ldots, f_t)$ where all intermediate states belong to $\mathcal{A}_t$. This determines a partition $(J_1, \ldots, J_n)$ of J such that if $j \in J_t$ then $w_j \in X_t$. The transitions of $\mathcal{A}$ ensure that i and f are the initial and final states of γ, respectively, and that i and f do not occur elsewhere in γ. Moreover the condition $P \supseteq \{i_1, \ldots, i_n, f_1, \ldots, f_n\}$ ensures that each J_i is dense in J.

In order to prove that $w \in \mathrm{sh}(X_1, \ldots, X_n)$, there remains to show that J is complete. Assume for a contradiction that J is not complete; this implies that J can be written as $J = J' + J''$ where J' has no greatest element and J'' has no least element. Thus we have $w = w' \cdot w''$ with $w' = \prod_{j \in J'} w_j$ and $w'' = \prod_{j \in J''} w_j$. We have $w' = (x_k)_{k \in K'}$ and $w'' = (x_k)_{k \in K''}$ with $K = K' + K''$, and K' has no greatest element and K'' has no least element. Therefore in the path γ, if k denotes the cut (K', K''), we must have $\gamma_k = q$ for some state q such that $\mathcal{A}$ admits transitions $P_1 \to q$ and $q \to P_2$, where P_1 is the limit set on the left of q and P_2 is the limit set on the right of q. But the very definitions of w' and w'' imply that $P_1, P_2 \supseteq \{i_1, \ldots, i_n, f_1, \ldots, f_n\}$, and the definition of transitions of $\mathcal{A}$ yields that $q \in \{i_1, \ldots, i_n, f\} \cap \{f_1, \ldots, f_n, i\}$ which is empty, leading to a contradiction.

The case where the empty word ε belongs to some X_t can be treated by a slight modification of the previous construction. Let $H = \{t \mid \varepsilon \in X_t\}$. Consider the automata $\mathcal{A}'$ obtained from $\mathcal{A}$ by adding a new state s_t for each $t \in H$, and by adding transitions of the form $P \to i_k, f, s_t$ and $f_k, i, s_t \to P$, for $k = 1, \ldots, n$, $t \in H$, and P satisfies

- for every $j \notin H$, $P \supseteq \{i_j, f_j\}$
- for every $j \in H$, $P \cap \{i_j, f_j, s_j\}$ equals either $\{i_j, f_j\}$, or $\{i_j, f_j, s_j\}$, or $\{s_j\}$

Let us show that $\mathcal{A}'$ accepts $\mathrm{sh}(X_1, \ldots, X_n)$. If $w \in \mathrm{sh}(X_1, \ldots, X_n)$, then one finds a successful path γ for w in $\mathcal{A}'$ in a similar way as in the previous case, except that for each w_j equal to ε and such that $j \in J_t$, we associate the path (s_t) to w_j. Conversely if $w = (x_k)_{k \in K}$ is a word indexed by the linear ordering K and is the label of a successful path $\gamma = (\gamma_k)_{k \in \hat{K}}$ in $\mathcal{A}$, then w can be written as $w = \prod_{j \in J} w_j$ where each w_j is either the label of a subpath of γ from i_t to f_t where all intermediate states belong to $\mathcal{A}_t$, or $w_j = \varepsilon$ is the label of some empty path from s_t to s_t. Note that the length of the word w is not complete but the ordering J is complete. Indeed, a gap in the length of w is labelled by some state s_t. For each of these cuts, a point j is added in J such that $w_j = \varepsilon$. These points make complete the ordering J. As in the previous case, this decompostion of w determines a partition $(J_1, \ldots, J_n)$ of J such that $w_j \in X_t$ whenever $j \in J_t$, and the very definition of $\mathcal{A}'$ ensures that J and $(J_1, \ldots, J_n)$ satisfy the required conditions.

7 From automata to rational expressions

In this section, we prove that for any automaton $\mathcal{A}$, there is a rational expression denoting the set of words accepted by $\mathcal{A}$. The proof is based on an extension of Yamada's classical approach.

The proof structure is similar to the one given in [5] for the case of languages of words indexed by countable scattered orderings. For the sake of readability we shall repeat some arguments of [5], and also point out the main differences with [5] along the proof.

We start with the following lemma, which gives a characterization of the orderings of the form $J \cup \hat{J}$. Lemma 8 of [5] is almost the same but the hypotheses on J and J' are weaker. They do not exclude a dense interval only containing elements of J'. This case cannot occur since the ordering K is scattered in [5].

LEMMA 11 *Let K be a complete linear ordering with a least and a greatest element, and let (J, J') be a partition of K. Suppose that any element of J has a predecessor and successor in J', and that there is at least one element of J between two elements of J'. Then J' equals $\hat{J}$, that is $K = J \cup \hat{J}$.*

Proof We define a function f from K into $J \cup \hat{J}$ as follows. For any $k \in K$, define

$$f(k) = \begin{cases} k & \text{if } k \in J \\ \big(\{j \in J \mid j < k\}, \{j \in J \mid k < j\}\big) & \text{if } k \in J'. \end{cases}$$

Since $J \cap J' = \varnothing$ and $K = J \cup J'$, the function f is well defined. The restriction of f to J is the identity. The image of an element of J' is a cut of J. Therefore f is a function from K into $J \cup \hat{J}$.

We claim that the function f is one-to-one. We first show that $k \neq k'$ implies $f(k) \neq f(k')$. If $k \in J$ or $k' \in J$, the result is trivial. Suppose then that $k, k' \in J'$ and that $k < k'$. By our second hypothesis there exists $j' \in J$ such that $k < j' < k'$, which implies $\{j \in J \mid j < k\} \neq (\{j \in J \mid j < k'\}$, thus $f(k) \neq f(k')$.

We now prove that the function f is onto. It is clear that $J \subseteq f(K)$. Let (L, M) be a cut of J. We claim that there is $k \in J'$ such that $(L, M) = f(k)$. Since K is complete and has a least and a greatest element, any subset of K has a greatest lower bound and a least upper bound. Define the two elements l and m of K by $l = \sup(L)$ and $m = \inf(M)$. If l belongs to L, it has a successor k in J' and one has $(L, M) = f(k)$. If m belongs to M, it has a predecessor k in J' and one has $(L, M) = f(k)$. If l and m do not belong to L and M, they belong to J' and their image by f is the cut (L, M). Since f is one-to-one, l and m are equal. □

Now let us prove that for any automaton $\mathcal{A}$, there is a rational expression denoting the set of words accepted by $\mathcal{A}$.

In the sequel we shall prove that every word accepted by $\mathcal{A}$ belongs to some rational language L. The task to check that every word $w \in L$ labels an accepting run of $\mathcal{A}$ is easy and is left to the reader.

We first introduce some notation. Let $\mathcal{A} = (Q, A, E, I, F)$ be a fixed automaton. The *content* $\mathrm{C}(\gamma)$ of a path γ is the set of states which occur inside γ. It does not take account the first and the last state of the path. Recall that a path γ labeled by a word of length J is a function from $\hat{J}$ into Q. The content of a path γ is thus formally defined by $\mathrm{C}(\gamma) = \gamma(\hat{J}^*)$. The *full content* $\mathrm{FC}(\gamma)$ of γ is defined by $\mathrm{FC}(\gamma) = \gamma(\hat{J})$.

A path γ from a state p to a state p' which is of content P and labeled by x is denoted by

$$\gamma : p \overset{x}{\underset{P}{\rightsquigarrow}} p'.$$

If $x \neq \varepsilon$, the path γ uses a first transition σ which leaves p and a last transition σ' which enters p'. To emphasize the use of σ and σ', the path γ is then denoted

$$\gamma : \sigma \overset{x}{\underset{P}{\rightsquigarrow}} \sigma'.$$

In both notations, we may omit the label or the content of the path if they are not relevant.

Let P be a subset of states and let σ and σ' be two transitions of $\mathcal{A}$. We define the sets of words $\Pi^P_{\sigma,\sigma'}$, $\nabla^P_{\sigma,\sigma'}$, $\Delta^P_{\sigma,\sigma'}$, and $\Gamma^P_{\sigma,\sigma'}$ as follows.

$$\begin{aligned}
\Pi^P_{\sigma,\sigma'} &= \{x \mid \sigma \overset{x}{\underset{P}{\rightsquigarrow}} \sigma'\} \\
\nabla^P_{\sigma,\sigma'} &= \{x \mid \sigma \overset{x}{\underset{P}{\rightsquigarrow}} \sigma' \text{ without any transition } P \to r\} \\
\Delta^P_{\sigma,\sigma'} &= \{x \mid \sigma \overset{x}{\underset{P}{\rightsquigarrow}} \sigma' \text{ without any transition } r \to P\} \\
\Gamma^P_{\sigma,\sigma'} &= \{x \mid \sigma \overset{x}{\underset{P}{\rightsquigarrow}} \sigma' \text{ without any transition } r \to P \text{ or } P \to r\}.
\end{aligned}$$

Note that *without any transition* $P \to r$ means that the path γ does not use any left limit transition of the form $P \to r$ for any $r \in \mathrm{C}(\gamma)$ except perhaps for the last transition if σ' is a left limit transition of this form. Thus, the left limit $\lim_{c^-} \gamma$ at any cut c different from the last cut must be a strict subset of P. Both sets $\nabla^P_{\sigma,\sigma'}$ and $\Delta^P_{\sigma,\sigma'}$ are subsets of $\Pi^P_{\sigma,\sigma'}$ and the set $\Gamma^P_{\sigma,\sigma'}$ is equal to the intersection $\nabla^P_{\sigma,\sigma'} \cap \Delta^P_{\sigma,\sigma'}$.

The paths considered in the definition of the sets $\Pi^P_{\sigma,\sigma'}$, $\nabla^P_{\sigma,\sigma'}$, $\Delta^P_{\sigma,\sigma'}$, and $\Gamma^P_{\sigma,\sigma'}$ use at least one transition. Therefore, the empty word is not contained in them. Since a path is successful if its first and last states are respectively initial and final, the set of words accepted by the automaton $\mathcal{A}$ is equal to the union

$$\varepsilon(\mathcal{A}) + \bigcup_{P \subseteq Q, \sigma \in \mathrm{Out}(I), \sigma' \in \mathrm{In}(F)} \Pi^P_{\sigma,\sigma'}$$

where $\varepsilon(\mathcal{A})$ is equal to ε if $I \cap F \neq \varnothing$ and to $\varnothing$ otherwise. We claim that any set $\Pi^P_{\sigma,\sigma'}$, $\nabla^P_{\sigma,\sigma'}$, $\Delta^P_{\sigma,\sigma'}$ and $\Gamma^P_{\sigma,\sigma'}$ is rational. The proof is by induction on the cardinality of P.

We first suppose that P is the empty set $\varnothing$. If both transitions σ and σ' are equal to the same successor transition $p \xrightarrow{a} q$, all four sets $\Pi^P_{\sigma,\sigma'}$, $\nabla^P_{\sigma,\sigma'}$, $\Delta^P_{\sigma,\sigma'}$ and $\Gamma^P_{\sigma,\sigma'}$ are equal to the singleton $\{a\}$. Otherwise, they are all empty. In both cases, there are rational. This completes the base case of the induction.

We now suppose that for any $R \subsetneq P$ and any transitions τ and τ', all four sets $\Pi^R_{\tau,\tau'}$, $\nabla^R_{\tau,\tau'}$, $\Delta^R_{\tau,\tau'}$ and $\Gamma^R_{\tau,\tau'}$ are rational. We claim that the sets $\Pi^P_{\sigma,\sigma'}$, $\nabla^P_{\sigma,\sigma'}$, $\Delta^P_{\sigma,\sigma'}$ and $\Gamma^P_{\sigma,\sigma'}$ can be expressed by rational expressions involving the sets $\Pi^R_{\tau,\tau'}$, $\nabla^R_{\tau,\tau'}$, $\Delta^R_{\tau,\tau'}$ and $\Gamma^R_{\tau,\tau'}$.

We refer the reader to the paper [5] for the proof that $\nabla^P_{\sigma,\sigma'}$, $\Delta^P_{\sigma,\sigma'}$ and $\Gamma^P_{\sigma,\sigma'}$ are rational, since this proof does not depend on any assumption on the linear ordering J. It is based on the ordering of the cuts where the left or right limit set is P. If only cuts where the left (or right) limit set is P occur, then the ordering of these cuts is well-ordered and thus scattered.

Therefore we only have to prove that $\Pi^P_{\sigma,\sigma'}$ is rational. For this case the arguments of [5] cannot still be used, as they heavily depend on the assumption that J is countable and scattered.

We shall use the following notions and properties. Let J be a linear ordering. A *condensation* of J is an equivalence relation $\sim$ on J such that each of its classes is an interval. Disjoint intervals are naturally ordered and the quotient $J/{\sim}$ can be endowed with a linear ordering induced by the ordering of J.

The quotient $J/{\sim}$ inherits some properties of J. If J has a least or a greatest element, then $J/\sim$ has also a least or a greatest element which are the class of the least or greatest element of J. If J is complete, then $J/{\sim}$ is also complete.

Let us mention some useful properties of consecutive classes k and k' of $J/{\sim}$ when the ordering J is complete. Define j and j' by $j = \sup(k)$ and $j' = \inf(k')$. By definition, one has $j \leq j'$. If $j < j'$, then j and j' respectively belong to k and k' since the classes k and k' are consecutive. In that case, the interval k is right closed and k' is left closed. If $j = j'$, then j belongs to either k or k'. In the former case, k is right closed and k' is left open and in the latter case, k is right open and k' is left closed. Note that it is impossible that k is right open and k' is left open.

In the sequel we will deal with intervals of $\hat{J}$. Each interval I has the form (c_1, c_2), $(c_1, c_2]$, $[c_1, c_2)$ or $[c_1, c_2]$, which will be expressed by saying that I has the *type* ( ), ( ], [ ) or [ ], respectively.

We shall consider successive condensations. Given a linear ordering J and a condensation $\sim_1$ of J, one can define another condensation $\sim_2$ over $K = J/{\sim}$. Then $\sim_2$ can be seen as a condensation over J, and in a similar way we shall often see equivalence classes of $\sim_2$ as intervals of J.

Now let us turn to the proof that $\Pi^P_{\sigma,\sigma'}$ is rational. We first give a sketch of the proof. It consists in cutting the path associated with any element of $\Pi^P_{\sigma,\sigma'}$ into intervals of the form ( ) or [ ] only, in such a way that the set of open intervals and the set of closed intervals interleave as in Lemma 11. This will allow to use the operation $\diamond$ to find an adequate rational expression. In the course of the proof the shuffle operation will also be involved.

The above collection of intervals ( ) and [ ] is defined via four successive condensations, that can be briefly described as follows. Let γ be a path labeled by a word $x = (a_j)_{j \in J}$ in $\Pi^P_{\sigma,\sigma'}$. The first condensation $\sim_1$ consists in cutting γ into intervals such that transitions of the form $P \to \ldots$ and $\ldots \to P$ in γ occur only at the extremity of these intervals. For each interval one can use the induction hypothesis to find a corresponding rational expression. The second condensation $\sim_2$ is obtained by merging elements of $K = J/{\sim_1}$ as follows: two elements $k_1 < k_2$ of K are merged iff all elements of the interval $[k_1, k_2]$, seen as intervals of $\hat{J}$, have the same type. The aim of the third condensation $\sim_3$ is to make intervals ( ] and [ ) "disappear". This is done by merging triplets of consecutive elements of $L = K/{\sim_2}$ which, seen as intervals of $\hat{J}$, have the form [ ),[ ],( ] (successively), and then merging the remaining couples of consecutive elements of L which, seen as intervals of $\hat{J}$, have the form [ ),[ ] (respectively [ ],( ]). This gives rise to a set $M = L/{\sim_3}$ of intervals of $\hat{J}$ that only have the form ( ) or [ ]. In the set M there can exist dense intervals consisting only in intervals of $\hat{J}$ of the form $[\,]$; the last condensation $\sim_4$ consists in an appropriate condensation of elements of M (the shuffle operation appears here) whose effect is to make "disappear" such dense intervals. The resulting quotient set $N = M/{\sim_4}$ consists only in intervals of $\hat{J}$ of the form $(\,)$ or $[\,]$, such that the set of intervals $(\,)$ and the set of intervals $[\,]$ interleave as Lemma 11 requires.

Let us point out that the two condensations $\sim_1$ and $\sim_2$ already appear in [5], and actually suffice to obtain a rational expression for $\Pi^P_{\sigma,\sigma'}$ if we consider only languages of words indexed by countable scattered orderings. Here we have to introduce two new condensations $\sim_3$ and $\sim_4$ to deal with all linear orderings.

As before, let γ be a path labeled by a word $x = (a_j)_{j \in J}$ in $\Pi^P_{\sigma,\sigma'}$, and let p and p' be its first and last state, respectively.

- **First condensation**

Consider the following condensation $\sim_1$ on $\hat{J}$ defined as follows. For any cuts $c_1, c_2 \in \hat{J}$, the relation

$$c_1 \sim_1 c_2$$

holds iff for any $c \in [c_1, c_2)$, $\lim_{c^+} \gamma \neq P$ and for any $c \in (c_1, c_2]$, $\lim_{c^-} \gamma \neq P$.

Any equivalence class of $\sim_1$ is an interval. Let us study the structure of the equivalence classes of $\sim_1$ and the consecutive elements of the quotient ordering $K = \hat{J}/{\sim_1}$. Note that K is a complete ordering with a least and a greatest element.

Let $k \in K$ be an equivalence class of $\sim_1$, with $c_1 = \inf(k)$ and $c_2 = \sup(k)$. For any $c \in (c_1, c_2)$, one has $\lim_{c^+} \gamma \neq P$ and $\lim_{c^-} \gamma \neq P$. By definition of $\sim_1$, one checks that k is left open iff $\lim_{c_1^+} \gamma = P$. Symmetrically, k is right open iff $\lim_{c_2^-} \gamma = P$.

Consider two consecutive elements k and k' of K. As $\hat{J}$ is complete, it is impossible that k is right open and k' is left open. We also claim that it is impossible that k is right closed and k' is left closed. Otherwise, the greatest element $\max(k)$ and the least element $\min(k')$ would be consecutive elements of $\hat{J}$, thus they would satisfy $\max(k) \sim_1 \min(k')$ and this is a contradiction. This proves that, either k is right open and k' is left closed, or k is right closed and k' is left open. Moreover, any element k of K which is right open always has a successor k' which is then left closed. Symmetrically, if k is left open, it has a predecessor k' which is right closed.

- **Second condensation**

We now define a condensation $\sim_2$ on the ordering K. For all $k_1, k_2 \in K$, the relation

$$k_1 \sim_2 k_1$$

holds iff elements of the interval $[k_1, k_2]$ have all the type $(\,]$, or have all the type $[\,)$.

As done for K, we study the quotient ordering $L = K/{\sim_2}$ which is a complete linear ordering with a least and a greatest element.

If l is an equivalence class of $\sim_2$ with elements of type $(\,)$, then l is a singleton since K is complete. In the same way, any equivalence class of $\sim_2$ with elements of type [ ] is a singleton. This is no longer true for the classes containing elements of type ( ] or [ ).

The ordering $\hat{J}$ is complete, thus any equivalence class l of $\sim_2$ that consists in elements of type [ ), always has a successor, and if l is left open, it also has a predecessor. In the same way, any class l with elements of type ( ] always has a predecessor and if l is right open, it also has a successor. Moreover, any class l reduced to a singleton of type ( ) has a predecessor and a successor.

Now consider two consecutive elements $l < l'$ of L. Since $\sim_2$ is a condensation, it is impossible that l is right open and l' is left open. Let $k = \sup(l)$ and $k' = \inf(l')$. Assume first that l and l' are right and left closed. Then k and k' are consecutive elements of K, for which we already know the possible configurations. Moreover, k and k' have different types since they are in different classes. Assume now that l is right open and that l' is left closed. If follows that $k = k'$ and the type of k is different from the type of the elements of l. As l is right open, two types are possible for its elements: ( ] or [ ). If the type is [ ), one checks that k has type [ ] due to the properties seen for K. If the type of the elements of l is ( ], then k has either type [ ] or [ ). The last case when l is right closed and l' is left open is symmetrical.

Let us go further. We consider L as a collection of intervals of $\hat{J}$ which partition $\hat{J}$, by composing the two condensations $\sim_1$ and $\sim_2$. To avoid any confusion, when L is seen as the quotient ordering over K, that is $L = K/{\sim_2}$, an equivalence class is described as an interval composed with elements k of K. When L is seen as the quotient ordering over $\hat{J}$, a class is described as an interval of elements c of $\hat{J}$. A class l is then seen as the interval $\bigcup_{k \in l} k$ of $\hat{J}$.

Let us detail the different cases. We begin with the classes l reduced to a singleton k of type ( ) or [ ]. Seen over $\hat{J}$, we respectively get $l = (c, c')$ or $l = [c, c']$. Let τ be the transition leaving $\gamma(c)$ and let τ' be the transition entering $\gamma(c')$. Therefore, if $l = (c, c')$, the label y of the path $\gamma(c) \rightsquigarrow \gamma(c')$ belongs to

$$\Gamma^P_{\tau,\tau'} \quad \text{with} \quad \tau = \gamma(c) \to P \text{ and } \tau' = P \to \gamma(c') \tag{1}$$

and if $l = [c, c']$, then either $y = \varepsilon$ (when $c = c'$) or y belongs to

$$Z_{\tau,\tau'} = \Gamma^P_{\tau,\tau'} \cup \bigcup_{R \subsetneq P} \Pi^R_{\tau,\tau'} \quad \text{with} \quad \tau \neq \gamma(c) \to P \text{ and } \tau' \neq P \to \gamma(c'). \tag{2}$$

We proceed with the classes l with all their elements of type [ ). Suppose that l is left open. We have seen before that over $\hat{J}$, l is an interval $l = (c, c')$ such that $\lim_{c^+} \gamma = P$ and $\lim_{c'^-} \gamma = P$. The path $\gamma(c) \rightsquigarrow \gamma(c')$ does not involve any right limit transition $r \to P$. Let τ be the transition leaving $\gamma(c)$ and let τ' be the transition entering $\gamma(c')$. Therefore, the label y of the path $\gamma(c) \rightsquigarrow \gamma(c')$ belongs to

$$\Delta^P_{\tau,\tau'} \quad \text{with} \quad \tau = \gamma(c) \to P \text{ and } \tau' = P \to \gamma(c'). \tag{3}$$

A similar description holds for classes l with elements of type ( ], which are right open. With the same notation, we have $l = (c, c')$ and the label y of the path $\gamma(c) \rightsquigarrow \gamma(c')$ belongs to

$$\nabla^P_{\tau,\tau'} \quad \text{with} \quad \tau = \gamma(c) \to P \text{ and } \tau' = P \to \gamma(c'). \tag{4}$$

Two cases have still to be considered: the case of left closed classes l with elements of type [ ) and the symmetrical case of right closed classes l with elements of type ( ]. In the former case, l is an interval $l = [c, c')$ over $\hat{J}$ such that $\lim_{c^+} \gamma \neq P$ and $\lim_{c'^-} \gamma = P$. The path $\gamma(c) \rightsquigarrow \gamma(c')$ is again without any right limit transition $r \to P$ and with the same notation as before, y belongs to

$$\Delta^P_{\tau,\tau'} \quad \text{with} \quad \tau \neq \gamma(c) \to P \text{ and } \tau' = P \to \gamma(c'). \tag{5}$$

In the latter case, a symmetric description holds with $l = (c, c']$ and y belongs to

$$\nabla^P_{\tau,\tau'} \quad \text{with} \quad \tau = \gamma(c) \to P \text{ and } \tau' \neq P \to \gamma(c'). \tag{6}$$

• **Third condensation**

We now define a third condensation $\sim_3$ on L, that we obtain by merging pairs or triples of consecutive elements of L as follows.

Consider an element $l = [c_1, c_2)$ of L as described in Equation (5). Recall that l has a successor l' which is necessarily a singleton k of type [ ]. Such a class l' has been described in Equation (2). The class l' is equal to an interval $l' = [c_2, c_3]$. Analogously, if $l'' = (c_3, c_4]$ is an element of L as described in Equation (6), it has a predecessor $l' = [c_2, c_3]$ as described in Equation (2).

Let $l' = [c_2, c_3]$ be an interval of Equation (2). If it has a predecessor $[c_1, c_2)$ of Equation (5) and a successor $(c_3, c_4]$ of Equation (6), then we merge the three intervals $[c_1, c_2)$, $[c_2, c_3]$ and $(c_3, c_4]$ in a single interval $[c_1, c_4]$. Otherwise, if l' has only a predecessor $[c_1, c_2)$ (respectively a successor $(c_3, c_4]$), then we merge the intervals $[c_1, c_2)$ and $[c_2, c_3]$ (respectively $[c_2, c_3]$ and $(c_3, c_4]$) in a single interval $[c_1, c_3]$ (resp. $[c_2, c_4]$).

Let $M = L/{\sim_3}$. Note that all elements of M, seen as intervals of $\hat{J}$, have the form ( ) and [ ] only. Let M_1 (resp. M_2) denote elements of M which are open (resp. closed) intervals of $\hat{J}$.

Consider $m = [c, c']$ in M_2. Let y be the label of path $\gamma(c) \rightsquigarrow \gamma(c')$ (see Equations (2), (5) and (6)). If $c = c'$ then $y = \varepsilon$. Assume now that $c \neq c'$; let τ be the transition leaving $\gamma(c)$ and let τ' be the transition entering $\gamma(c')$. The label y belongs to

$$Y_{\tau,\tau'} = Y'_{\tau,\tau'} \cup Z_{\tau,\tau'} \tag{7}$$

where $Y'_{\tau,\tau'}$ corresponds to the case where m is obtained by merging two or three consecutive elements of L, and $Z_{\tau,\tau'}$ corresponds to the case where m is an interval of $\hat{J}$ which comes directly from the first condensation (i.e. m corresponds to an equivalence class of $\sim_1$ which was not merged with other classes during the second and third condensations).

We have

$$\begin{aligned} Y'_{\tau,\tau'} = & \bigcup_{\substack{(\tau'_1,\tau_1)\in\mathcal{T}_1 \\ (\tau'_2,\tau_2)\in\mathcal{T}_2}} \Delta^P_{\tau,\tau'_1} Z_{\tau_1,\tau'_2} \nabla^P_{\tau_2,\tau'} \quad \cup \quad \bigcup_{(\tau'_3,\tau_3)\in\mathcal{T}_3} \Delta^P_{\tau,\tau'_3} \nabla^P_{\tau_3,\tau'} \\ & \cup \bigcup_{(\tau'_1,\tau_1)\in\mathcal{T}_1} \Delta^P_{\tau,\tau'_1} Z_{\tau_1,\tau'} \quad \cup \quad \Delta^P_{\tau,\tau'} \\ & \cup \bigcup_{(\tau'_2,\tau_2)\in\mathcal{T}_2} Z_{\tau,\tau'_2} \nabla^P_{\tau_2,\tau'} \quad \cup \quad \nabla^P_{\tau,\tau'} \end{aligned} \tag{8}$$

where the sets $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ are defined by

$$\begin{aligned} \mathcal{T}_1 &= \{(\tau'_1,\tau_1) \mid \exists q \ \ \tau'_1 = P \to q, \tau_1 \in \text{Out}(q) \text{ and } \tau_1 \neq q \to P\} \\ \mathcal{T}_2 &= \{(\tau'_2,\tau_2) \mid \exists q \ \ \tau'_2 \in \text{In}(q), \tau'_2 \neq P \to q \text{ and } \tau_2 = q \to P\} \\ \mathcal{T}_3 &= \{(\tau'_3,\tau_3) \mid \exists q \ \ \tau'_3 = P \to q \text{ and } \tau_3 = q \to P\}. \end{aligned}$$

Let us give some details about Equation (8):

- the first line corresponds to the case where three consecutive elements of L (of type $[\,)$, $[\,]$, $(\,]$, respectively) are merged. The second term of the union allows to deal with the case where the element of type $[\,]$ is reduced to a singleton.

- the second line corresponds to the case where two consecutive elements of L (of type $[\,)$, $[\,]$, respectively) are merged. As before, the term $\Delta^P_{\tau,\tau'}$ allows to deal with the case where the element of type $[\,]$ is a singleton.

- the third line corresponds to the case where two consecutive elements of L (of type $[\,]$, $(\,]$, respectively) are merged. The last term allows to deal with the case where the element of type $[\,]$ is a singleton.

Note that if m is not the first element of $M_1 \cup M_2$, then τ belongs to $T_1 = \{\tau_1 \mid \exists\tau'_1 \ \ (\tau'_1,\tau_1) \in \mathcal{T}_1\}$ and if m is not the last element of $M_1 \cup M_2$, then τ' belongs to $T'_2 = \{\tau'_2 \mid \exists\tau_2 \ \ (\tau'_2,\tau_2) \in \mathcal{T}_2\}$. Otherwise, the set $Y_{\tau,\tau'}$ is such that $\tau = \sigma$ or $\tau' = \sigma'$. Moreover, when $\tau = \sigma$ and $\tau' = \sigma'$, the definition of $Z_{\tau,\tau'}$ given by Equation (2) must be slightly changed into $Z_{\sigma,\sigma'} = \Gamma^P_{\sigma,\sigma'}$ due to the content equal to P.

• Fourth condensation

At this step there could exist in M dense intervals consisting only in elements of M_2. Indeed assume that there exist $m_1, m_2 \in M_1$ such that $m_1 < m_2$, and $]m_1, m_2[\cap M_1 = \varnothing$. Since there cannot exist in M two consecutive elements that belong to M_2 (by the very definition of $\sim_3$), the interval $]m_1, m_2[$ of M consists either in a singleton $m \in M_2$, or in a dense interval of elements of M_2.

We shall define a (last) condensation $\sim_4$ which will merge elements of such dense intervals of elements of M_2. This will involve the shuffle operation for the corresponding rational expression.

Given a linear ordering $(J,<)$ and a set F, we call *labeling of J by F* any function $l : J \to F$. We say that an interval I of J is *homogeneous* for l if for every $t \in l(I)$, the set $\{x \in I \mid l(x) = t\}$ is dense in I.

LEMMA 12 *If J is a dense infinite linear ordering and F is a finite set, and $l : J \to F$ is a labeling of J by F, then there exists a non-trivial interval I of J which is homogeneous for l.*

Proof Easy induction on the cardinality of F. □

Consider the finite labeling $l : M \to 2^P$ which maps every element $m \in M$ to the *full content* of m (when m is seen as an interval of $\hat{J}$). The condensation $\sim_4$ on M is obtained by merging, within every interval $]m_1, m_2[$ of M such that $m_1, m_2 \in M_1$ and $]m_1, m_2[\cap M_1 = \varnothing$, all elements of M that belong to the same homogeneous interval with respect to the labeling l. More formally, for $m, m' \in M$ such that $m < m'$ we set $m \sim_4 m'$ iff $[m, m']$ is contained in an open interval I such that $I \cap M_1 = \varnothing$ and I is homogeneous for l.

Observe that this merges elements of M_2, which are closed intervals of $\hat{J}$, into open intervals of $\hat{J}$.

Let $N = M/{\sim_4}$. By the very definition of $\sim_4$, N consists only in closed and open intervals of $\hat{J}$. Let us denote by N_1 (respectively N_2) the set of elements of N which are open (respectively closed) intervals of $\hat{J}$. We denote by N_1' the elements of N_1 which appeared during the fourth condensation, i.e. which were obtained by merging a dense interval of elements of M_2.

LEMMA 13 *The ordering N and the partition (N_1, N_2) of N satisfy the hypotheses of Lemma 11. Thus $N_2 = \hat{N}_1$, that is $N = N_1 \cup \hat{N}_1$.*

Proof First of all, N is a condensation of M, thus N is complete with a least and a greatest element. Now consider an element n of N_1. Since n is open it has predecessor and successor which are closed and thus in N_2.

It remains to show that for all elements $n, n' \in N_2$ with $n < n'$, there exists $n'' \in N_1$ such that $n < n'' < n'$. Assume for a contradiction that there is no such n''. Observe first that n, n' cannot be consecutive elements of N. Indeed n and n' are closed intervals of $\hat{J}$, and since $\sim_1$ is a refinement of $\sim_4$, this would imply the existence of two consecutive equivalence classes k_1, k_2 of $\sim_1$ such that k_1 is right closed and k_2 is left closed, which was shown to be impossible. The fact that n, n' are not consecutive, and that the interval $[n, n']$ of N contains only elements of N_2, yield that $[n, n']$ is a dense interval of elements of N_2. Now every element of N_2 coincides, as an interval of $\hat{J}$, with an equivalence class of $\sim_3$ (in M_2), since elements of N_2 are closed intervals of $\hat{J}$ and the condensation $\sim_4$ only merges classes of $\sim_3$ which are closed intervals of $\hat{J}$, into open intervals. Therefore we can see the interval $[n, n']$ as an interval of elements of M_2, and apply Lemma 12 to get a non-trivial interval I of $[n, n']$ which is homogeneous for our labeling l (restricted to $[n, n']$). From the definition of $\sim_4$ it follows that $[n, n']$, seen as an interval of N, contains an element of N_1', which contradicts our hypothesis. □

We shall give now rational expressions that correspond to elements of N (seen as intervals of $\hat{J}$).

Let us first give a rational expression for open intervals of $\hat{J}$ which correspond to elements of N_1'. Let $z = (c, c')$ in N_1'. By the very definition of $\sim_4$, z was obtained by merging a dense interval I of elements of M_2, where I is homogeneous for the labeling l. Let $R_1, \ldots, R_n \subseteq P$ be the the distinct values $l(m)$ for $m \in I$. We have $\mathrm{FC}(z) = \cup_{1 \leq i \leq n} R_i$. On the other hand z contains infinitely many equivalence classes of $\sim_3$ since it belongs to N_1', which implies $\mathrm{FC}(z) = P$. We have shown that $\cup_{1 \leq i \leq n} R_i = P$.

¿From the above arguments it follows that the label of the path $\gamma(c) \rightsquigarrow \gamma(c')$ belongs to

$$U_P = \bigcup_{n \geq 1, R_i \neq R_j, \cup_{i=1}^{n} R_i = P} \mathrm{sh}(S_{R_1}, \ldots, S_{R_n}). \tag{9}$$

where, for every $R \subseteq P$, S_R denotes the language of labels of paths $\gamma(c_1) \rightsquigarrow \gamma(c_2)$ having a full content equal to R, and such that $(c_1, c_2) \in M_2$, and $P \to \gamma(c_2)$ and $\gamma(c_1) \to P$ are transitions of $\mathcal{A}$.

For every $R \subsetneq P$ one has

$$S_R = \bigcup_{(\tau, \tau') \in \mathcal{T}_R} \Pi_{\tau, \tau'}^{R'}$$

where $\mathcal{T}_R$ denotes the set of couples (τ, τ') such that $\tau \in \mathrm{Out}(q)$, $\tau' \in \mathrm{In}(q')$, $R = R' \cup \{q, q'\}$, and $P \to q$ and $q' \to P$ are transitions of $\mathcal{A}$.

Further S_P satisfies

$$S_P = \bigcup_{(\tau, \tau') \in \mathcal{T}_P} \Pi_{\tau, \tau'}^{R'} \cup \bigcup_{(\tau, \tau') \in \mathcal{T}'} Y_{\tau, \tau'}'$$

where

- $\mathcal{T}_P$ denotes the set of couples (τ, τ') such that $\tau \in \mathrm{Out}\, q$, $\tau' \in \mathrm{In}(q')$, $P = R' \cup \{q, q'\}$, and $P \to q$ and $q' \to P$ are transitions of $\mathcal{A}$.
- $\mathcal{T}'$ denotes the set of couples (τ, τ') such that $\tau \in \mathrm{Out}\, q$, $\tau' \in \mathrm{In}(q')$, and $P \to q$ and $q' \to P$ are transitions of $\mathcal{A}$.

This shows that U_P is rational.

We can give now a rational expression for open intervals of $\hat{J}$ which correspond to elements of N_1. Let $z = (c, c')$ in N_1, the label of the path $\gamma(c) \rightsquigarrow \gamma(c')$ belongs to (see Equations (1), (3) and (4) and (9))

$$X_{\tau, \tau'} = \Gamma_{\tau, \tau'}^{P} \cup \Delta_{\tau, \tau'}^{P} \cup \nabla_{\tau, \tau'}^{P} \cup U_P \quad \text{with} \quad \tau = \gamma(c) \to P \text{ and } \tau' = P \to \gamma(c').$$

Consider now elements of N_2. It follows from the very definition of $\sim_4$ that the rational expression which corresponds to elements of N_2 (seen as closed intervals of $\hat{J}$) is the same as the one for elements of M_2. More precisely let $x =$

$[c,c']$ in N_2. Assume first that $c \neq c'$. Let y be the label of path $\gamma(c) \rightsquigarrow \gamma(c')$, and let τ be the transition leaving $\gamma(c)$ and let τ' be the transition entering $\gamma(c')$. Then the label y belongs to the language $Y_{\tau,\tau'}$ as defined in Equation (7).

Let us consider the case $c = c'$. If $\hat{N}_1^*$ contains an element $[c,c]$, then $N_1 \neq \varnothing$ and there are in $\mathcal{A}$ two transitions $P \to q$ and $q \to P$ for some state q. Analogously, if the first element of $\hat{N}_1$ is equal to $[c,c]$, then $N_1 \neq \varnothing$ and the first transition σ is equal to $p \to P$ (similarly for the last element of $\hat{N}_1$).

Finally, we consider the label $x \in \Pi^P_{\sigma,\sigma'}$ of the path γ. We shall use the languages $Y_{\tau,\tau'}$ as defined in Equation (7). If $N_1 = \varnothing$, then $\hat{N}_1$ is reduced to a singleton $[c,c']$ with $c \neq c'$. It follows that $x \in Y_{\sigma,\sigma'}$. If $N_1 \neq \varnothing$, we decompose x thanks to the rational operation $\diamond$ used on $N_1 \cup \hat{N_1}^*$. Different cases have to be considered depending on the transitions σ and σ'. Define the two sets X and Y by

$$X = \bigcup_{\substack{\tau = q \to P \\ \tau' = P \to q'}} X_{\tau,\tau'} \qquad \text{and} \qquad Y = \bigcup_{\substack{\tau \in T_1 \\ \tau' \in T_2'}} Y_{\tau,\tau'}.$$

Add to Y the empty word if there are in $\mathcal{A}$ two transitions $P \to q$ and $q \to P$ for some state q. Define the set $Y_\sigma = \bigcup_{\tau' \in T_2'} Y_{\sigma,\tau'}$ if $\sigma \neq p \to P$ and $Y_\sigma = \varepsilon$ if $\sigma = p \to P$. Define also the set $Y_{\sigma'} = \bigcup_{\tau \in T_1} Y_{\tau,\sigma'}$ if $\sigma' \neq P \to p'$ and $Y_{\sigma'} = \varepsilon$ if $\sigma' = P \to p'$. Then the label x belongs to $Y_\sigma (X \diamond Y) Y_{\sigma'}$ showing the inclusion

$$\Pi^P_{\sigma,\sigma'} \subseteq Y_{\sigma,\sigma'} \cup Y_\sigma (X \diamond Y) Y_{\sigma'}.$$

It can be verified that the other inclusion holds. Clearly, the set $Y_{\sigma,\sigma'}$ is included in $\Pi^P_{\sigma,\sigma'}$. One checks that the right limit transitions $q \to P$ and the left limit transitions $P \to q'$ involved in the operation $\diamond$ are well managed by the conditions imposed by T_1 and T_2'.

Therefore, $\Pi^P_{\sigma,\sigma'}$ is expressed as a rational expression on the sets $\nabla^P_{\tau,\tau'}$, $\Delta^P_{\tau,\tau'}$, $\Gamma^P_{\tau,\tau'}$ and $\Pi^R_{\tau,\tau'}$ with $R \subsetneq P$. This completes the proof.

8 Conclusion and open questions

We considered rational expressions and automata for words indexed by linear orderings, and prove that these two formalisms capture the same languages.

A natural question is whether the class of recognizable languages is closed under complementation. It has been proved in [21] that the answer is positive when one considers only words indexed by countable and scattered orderings (the proof relies upon semigroup theory). However the answer is negative in the general case: one can prove that the set of words indexed by a non scattered ordering is recognizable, while its complement is not.

The connections between automata over linear orderings and logic would be interesting to explore. In his seminal paper [6], Büchi proved that recognizable languages of finite words coincide with languages definable in the weak monadic second order theory of $(\omega, <)$, which allowed him to prove decidability of this theory. In [7] he proved that a similar equivalence holds between recognizable languages of infinite words of length ω and languages definable in the monadic second order theory of $(\omega, <)$. The result was then extended to languages of words indexed by a countable ordinal [8]. What can be said about monadic second order theories for linear orderings beyond ordinals ? Using the automata technique, Rabin proved in [19] decidability of the monadic second order theory of the binary tree, from which he deduces decidability of the monadic second order theory of $\mathbb{Q}$, which in turn implies decidability of the monadic second order theory of countable linear orderings. On the other hand, Shelah [23] improved model-theoretical techniques [25] that allow him to reprove almost all known decidability results about monadic second order theories, as well as new decidability results for the case of linear orderings. On the other hand he proved that the monadic second order theory of the real line is undecidable. Shelah's decidability method is model-theoretical, and up to now no corresponding automata techniques are known. This led Thomas to ask [25] whether there is an appropriate notion of automata for words indexed by linear orderings beyond the ordinals. As mentioned in [3], this question was an important motivation for the introduction of automata considered in the present paper. It would be interesting to provide a logical characterization of recognizable languages of words over linear orderings.

References

[1] E. Asarin, P. Caspi, and O. Maler. A Kleene theorem for timed automata. In *Proceedings, Twelfth Annual IEEE Symposium on Logic in Computer Science*, pages 160–171, 1997.

[2] B. Bérard and C. Picaronny. Accepting Zeno words without making time stand still. In *Mathematical Foundations of Computer Science 1997*, volume 1295 of *Lect. Notes in Comput. Sci.*, pages 149–158, 1997.

[3] V. Bruyère and O. Carton. Automata on linear orderings. In J. Sgall, A. Pultr, and P. Kolman, editors, *MFCS'2001*, volume 2136 of *Lect. Notes in Comput. Sci.*, pages 236–247, 2001.

[4] V. Bruyère and O. Carton. Hierarchy among automata on linear orderings. In R. Baeza-Yate, U. Montanari, and N. Santoro, editors, *Foundation of Information technology in the era of network and mobile computing*, pages 107–118. Kluwer Academic Publishers, 2002.

[5] Véronique Bruyère and Olivier Carton. Automata on linear orderings. Technical Report 2000–12, Institut Gaspard Monge, 2000.

[6] J. R. Büchi. Weak second-order arithmetic and finite automata. *Z. Math. Logik und grundl. Math.*, 6:66–92, 1960.

[7] J. R. Büchi. On a decision method in the restricted second-order arithmetic. In *Proc. Int. Congress Logic, Methodology and Philosophy of science, Berkeley 1960*, pages 1–11. Stanford University Press, 1962.

[8] J. R. Büchi. Transfinite automata recursions and weak second order theory of ordinals. In *Proc. Int. Congress Logic, Methodology, and Philosophy of Science, Jerusalem 1964*, pages 2–23. North Holland, 1965.

[9] O. Carton. Accessibility in automata on scattered linear orderings. In K.Diks and W.Rytter, editors, *MFCS'2002*, volume 2420 of *Lect. Notes in Comput. Sci.*, pages 155–164, 2002.

[10] Y. Choueka. Finite automata, definable sets, and regular expressions over ω^n-tapes. *J. Comput. System Sci.*, 17(1):81–97, 1978.

[11] B. Courcelle. Frontiers of infinite trees. *RAIRO Theoretical informatics*, 12(4):319–337, 1978.

[12] C. Dima. Real-time automata. *Journal of Automata, Languages and Combinatorics*, 6(1):3–24, 2001.

[13] D. Girault-Beauquier. Bilimites de langages reconnaissables. *Theoret. Comput. Sci.*, 33(2–3):335–342, 1984.

[14] M. R. Hansen, P. K. Pandya, and Z. Chaochen. Finite divergence. *Theoret. Comput. Sci.*, 138(1):113–139, 1995.

[15] S. Heilbrunner. An algorithm for the solution of fixed-point equations for infinite words. *RAIRO Theoretical informatics*, 14(2):131–141, 1980.

[16] S. C. Kleene. Representation of events in nerve nets and finite automata. In C.E. Shannon, editor, *Automata studies*, pages 3–41. Princeton university Press, Princeton, 1956.

[17] D. E. Muller. Infinite sequences and finite machines. In *Switching Circuit Theory and Logical Design: Proc. Fourth Annual Symp.*, pages 3–16. I.E.E.E., New York, 1963.

[18] M. Nivat and D. Perrin. Ensembles reconnaissables de mots bi-infinis. In *Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing*, pages 47–59, 1982.

[19] M.O. Rabin. Decidability of second-order theories and automata on infinite trees. *Transactions of the American Mathematical Society*, 141:1–35, 1969.

[20] A. Rabinovich. Star free expressions over the reals. *Theoret. Comput. Sci.*, 233(1–2):233–245, 2000.

[21] C. Rispal and O. Carton. Complementation of rational sets on countable scattered linear orderings. In C. S. Calude, E. Calude, and M. J. Dinneen, editors, *DLT'2004*, volume 3340 of *Lect. Notes in Comput. Sci.*, pages 381–392, 2004.

[22] J. G. Rosenstein. *Linear orderings.* Academic Press, New York, 1982.

[23] S. Shelah. The monadic theory of order. *Annals of Mathematics*, 102:379–419, 1975.

[24] W. Thomas. On frontiers of regular sets. *RAIRO Theoretical informatics*, 20:371–381, 1986.

[25] W. Thomas. Ehrenfeucht games, the composition method, and the monadic theory of ordinal words. In *Structures in Logic and Computer Science, A Selection of Essays in Honor of A. Ehrenfeucht*, number 1261 in Lect. Notes in Comput. Sci., pages 118–143. Springer-Verlag, 1997.

[26] O. Carton V. Bruyère and G. Sénizergues. Tree automata and automata on linear orderings. In T. Harju and J. Karhumäki, editors, *WORDS'2003*, pages 222–231. Turku Center for Computer Science, 2003.

[27] J. Wojciechowski. Finite automata on transfinite sequences and regular expressions. *Fundamenta informaticæ*, 8(3-4):379–396, 1985.