

# Constructible Words Characterize Rational Languages of Words Indexed by Scattered Linear Orderings

Thomas Braipson ✉ 🏠 

Montefiore Institute, University of Liège, Belgium

Tom Clara ✉ 🏠 

Montefiore Institute, University of Liège, Belgium

---

## Abstract

Automata on linear orderings are finite-state automata introduced by Bruyère and Carton as a broad generalization of finite, infinite and transfinite-word automata. In this context, a word is defined as a function from a linear ordering to a finite alphabet. This general definition can make automata on linear orderings difficult to reason about. In this work, we introduce constructible words as an intuitive way of tackling this difficulty. These words can be obtained by a finite number of applications of simple operators and thus admit a finite notation. We show that a rational language of words indexed by scattered (countable and uncountable) linear orderings is characterized by its constructible words. Our proof of this result relies on an interesting theorem of semigroup theory due to Colcombet. We expect this property to be useful in future theoretical developments about automata on scattered linear orderings.

**2012 ACM Subject Classification** Theory of computation → Automata over infinite objects

**Keywords and phrases** Automata on linear orderings, Rational languages, Ultimately periodic words, Constructible Words, Complementation, Algebraic properties of automata, Semigroups

**Digital Object Identifier** 10.4230/LIPIcs.MFCS.2026.25

**Related Version** *Full Version*: <https://doi.org/10.48550/arXiv.2607.01858>

**Funding** This work is partially supported by the FNRS-DFG PDR Weave (SMT-ART) grant 40019202.

*Thomas Braipson*: has been granted with a FRIA grant.

*Tom Clara*: is a Research Fellow of the F.R.S.-FNRS.

**Acknowledgements** The authors wish to thank Olivier Carton for discussing with them and advertising Colcombet's theorem. They also wish to thank the anonymous reviewers for their comments and ideas to go beyond this paper.

## 1 Introduction

Automata on linear orderings [5, 3, 4] were introduced by Bruyère and Carton and consist of a broad and elegant generalization of finite-word automata, infinite-word automata [23], automata on bi-infinite words [16] and transfinite words [24, 11]. In the context of automata on linear orderings, a word is defined as a function from any linear (i.e., total) ordering to a finite alphabet. This flexible definition generalizes the previously mentioned common notions of words.

Reasoning about automata on linear orderings (for instance, showing that two automata accept exactly the same words) can be tedious, since some complex words do not admit a natural mental representation. In this work, we introduce a theoretical tool that addresses this problem. We first define a *constructible* word as a word that admits a canonical, finite representation. We then prove that automata on scattered (countable and uncountable)



© Thomas Braipson and Tom Clara;

licensed under Creative Commons License CC-BY 4.0

51st International Symposium on Mathematical Foundations of Computer Science (MFCS 2026).

Editors: Michal Koucký and Daniela Petrişan; Article No. 25; pp. 25:1–25:18

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

linear orderings that accept the same constructible words are equivalent. Stated differently, rational languages of words indexed by scattered linear orderings are characterized by their constructible words. We expect this result to be useful in further developments.

Our work extends the characterization of  $\omega$ -regular languages by ultimately periodic words [7, Fact 1]. A simple proof of this classical result relies on the closure of  $\omega$ -regular languages under complementation [6]. However, rational languages of words indexed by scattered (countable and uncountable) linear orderings are not closed under complementation [18], hence our proof relies on different principles.

Sketchily, our proof is based on the fact that some words cannot be distinguished by a finite-state automaton, which naturally induces an equivalence relation on words. This idea, already used by Büchi in [6], was then adapted by Carton and Rispal to prove that rational languages on finite-rank scattered linear orderings are closed under complementation [10]. We reuse this idea in the case of words on scattered linear orderings, and we show that each equivalence class induced by an automaton contains a constructible word.

Much work has already been carried out regarding rational languages on *countable* scattered linear orderings. These languages were shown to coincide with those recognizable by a special form of semigroup [18, Theorem 14]. As a consequence, they are closed under complementation [18, Theorem 16], and definable in the monadic second-order theory of order (MSO) [2]. Carton also showed that any non-empty rational language on scattered linear orderings contains a countable constructible word [8, Theorem 2]. In the case of countable scattered words, our result follows immediately from these previous works. It can also be deduced from several results about the decidability [20] and expressivity [9] of MSO on countable domains. Hence our main contribution is in fact to provide a characterization of rational languages that goes beyond countability.

Our proofs exploit connections between rational languages and semigroups. It turns out that the algebraic properties of rational languages on unrestricted (i.e., countable and uncountable) scattered linear orderings are not as strong as they are in the countable case. Nevertheless, they remain useful, especially when combined with powerful semigroup results such as the factorization forests theorem due to Colcombet [12, Theorem 4], which is a generalization of Simon's factorization forests theorem [22]. Our work strongly relies on this result, and provides an example of use of its infinite version.

The remainder of this article is organized as follows. Section 2 recalls some notions related to automata on linear orderings. Then, in Section 3 we formally define the notion of *constructible* word. Section 4 introduces an equivalence relation on words and then motivates the link between automata and semigroups. Finally, in Section 5, we present Colcombet's theorem and apply it to show that two automata on scattered linear orderings are equivalent if and only if they accept the same constructible words.

## 2 Preliminary Notions

### 2.1 Orderings and Cuts

In this section, we recall some definitions on linear orderings. Further details can be found in [19]. A *linear ordering* is a set  $J$  equipped with a binary relation  $<_J$  that is

- total (for every  $j \neq k \in J$ , either  $j <_J k$  or  $k <_J j$ ),
- irreflexive (for every  $j \in J$ ,  $j \not<_J j$ ),
- antisymmetric (for every  $j, k \in J$ ,  $j <_J k$  implies  $k \not<_J j$ ),
- and transitive (for every  $j, k, \ell \in J$ ,  $j <_J k$  and  $k <_J \ell$  imply  $j <_J \ell$ ).

We simply write  $<$  in place of  $<_J$  when the ordering  $J$  is clear from the context.

Given a linear ordering  $J$ , a subordering of  $J$  is a subset  $I \subseteq J$  equipped with the order relation  $<_I$  defined as the restriction of  $<_J$  to  $I$ . A linear ordering  $J$  is *dense* if for any  $j < k \in J$ , there exists  $\ell \in J$  such that  $j < \ell < k$ . A linear ordering is *scattered* if it does not contain any infinite dense subordering<sup>1</sup>. This work focuses on scattered (countable and uncountable) linear orderings.

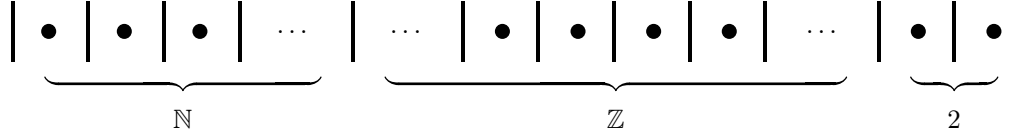
Two elements  $j < k \in J$  are said to be *consecutive* if there does not exist  $\ell \in J$  such that  $j < \ell < k$ . In such a case,  $k$  is the *successor* of  $j$ , and  $j$  the *predecessor* of  $k$ .

A *cut* of a linear ordering  $J$  is an ordered pair  $(K, L)$  such that  $\{K, L\}$  is a partition of  $J$ , and for every  $k \in K$  and  $\ell \in L$ , one has  $k < \ell$ . The set of cuts of  $J$  is denoted by  $\hat{J}$ . This set is itself a linear ordering under the relation  $<_{\hat{J}}$  such that  $(K_1, L_1) <_{\hat{J}} (K_2, L_2)$  if and only if  $K_1 \subsetneq K_2$  (or equivalently  $L_2 \subsetneq L_1$ ). The linear ordering  $\hat{J}$  has a least and a greatest element. These are respectively called the *first cut*  $\hat{J}_{\min} = (\emptyset, J)$  and the *last cut*  $\hat{J}_{\max} = (J, \emptyset)$ . The set  $\hat{J} \setminus \{\hat{J}_{\min}, \hat{J}_{\max}\}$  of *non-extremal* cuts of  $J$  is denoted by  $\hat{J}^*$ .

Given two linear orderings  $J_1$  and  $J_2$ , the ordering  $J = J_1 + J_2$  is defined as the ordering obtained by placing the elements of  $J_1$  before those of  $J_2$ , both sets keeping their internal order among their elements. Formally,  $J$  can be defined as the set  $(J_1 \times \{1\}) \cup (J_2 \times \{2\})$  such that  $(j, m) < (k, n)$  if and only if either  $m <_{\mathbb{N}} n$ , or  $m = n$  and  $j <_{J_m} k$ . More generally, given the linear orderings  $J$  and  $K_j$  for each  $j \in J$ , the ordering  $\sum_{j \in J} K_j$  is obtained, schematically, by replacing every element  $j$  of  $J$  by the corresponding set  $K_j$ , keeping the internal order among the elements of this set. Formally, we have  $\sum_{j \in J} K_j = \{(k, j) \mid j \in J \wedge k \in K_j\}$ , with  $(k_1, j_1) < (k_2, j_2)$  if and only if either  $j_1 <_J j_2$ , or  $j_1 = j_2$  and  $k_1 <_{K_{j_1}} k_2$ .

Given a linear ordering  $J$ , the *reverse* linear ordering  $-J$  is obtained by keeping the same elements as  $J$  and by flipping the order relation  $<_J$ . For two elements  $i, j \in -J$ , we thus have  $i <_{-J} j$  if and only if  $j <_J i$ .

► **Example 1.** The linear ordering  $\mathbb{N} + \mathbb{Z} + 2$  and its cuts are depicted in Figure 1, where dots represent the elements of  $\mathbb{N} + \mathbb{Z} + 2$  and vertical bars represent its cuts.



■ **Figure 1** The linear ordering  $\mathbb{N} + \mathbb{Z} + 2$  and its cuts.

## 2.2 Words Indexed by Linear Orderings

This section introduces a few definitions taken from [5]. A *word indexed by a linear ordering* is a function  $w : J \rightarrow \Sigma$  from a linear ordering  $J$  to a finite alphabet  $\Sigma$ . We say that  $J$  is the *length* of  $w$ , which is denoted by  $J = |w|$ , and that  $\hat{J}$  are its cuts. In the particular case where  $J$  is finite, the word  $w$  is simply a finite word. The notions of infinite and bi-infinite words correspond to the cases where  $J = \mathbb{N}$  and  $J = \mathbb{Z}$ , respectively. The empty word is the only word indexed by the empty linear ordering. In the remainder of this article, we will consider words indexed by arbitrary scattered (countable and uncountable) linear orderings, that we will call *scattered words*. The class of all the scattered words on the alphabet  $\Sigma$  is noted  $\Sigma^{\text{scat}}$ .

<sup>1</sup> Notice that according to the definition of density, a singleton is a dense linear ordering. Hence there exist dense linear orderings that are not infinite.

In the context of automata on linear orderings, two words  $w_1$  and  $w_2$  are said to be *isomorphic* if there exists a bijection  $\tau : |w_1| \rightarrow |w_2|$  that preserves both the symbols and the order: for all  $j, k \in |w_1|$ , we must have  $w_1(j) = w_2(\tau(j))$  and  $j <_{|w_1|} k \implies \tau(j) <_{|w_2|} \tau(k)$ . In the remainder of this article, we do not distinguish isomorphic words, since the automata that we will introduce in Section 2.4 are not able to discriminate them. Note that the length of a word should therefore be an *order type* rather than an ordering (see [5] for more details). In this work, we do not distinguish these two notions and only use the term “linear ordering”, even though it is a slight abuse of language.

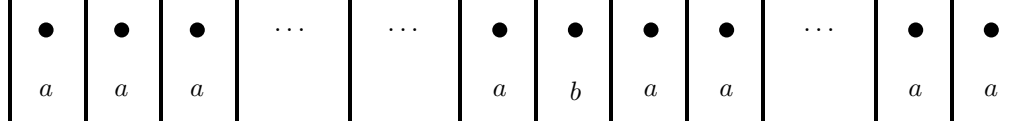
Given two words  $w_1 : K_1 \rightarrow \Sigma$  and  $w_2 : K_2 \rightarrow \Sigma$  (where  $K_1$  and  $K_2$  are assumed disjoint), the *concatenation* of  $w_1$  and  $w_2$ , denoted by  $w_1 \cdot w_2$  (or  $w_1 w_2$ ), is the word  $w : K_1 + K_2 \rightarrow \Sigma$  defined as

$$w((k, n)) = \begin{cases} w_1(k) & \text{if } n = 1 \\ w_2(k) & \text{if } n = 2. \end{cases}$$

The concatenation can be considered as a particular case of a more general operation. Given a linear ordering  $J$  and a set of words  $\{w_j : K_j \rightarrow \Sigma \mid j \in J\}$ , the concatenation of these words, denoted by  $\prod_{j \in J} w_j$ , is a word  $w$  indexed by  $\sum_{j \in J} K_j$ , and such that  $w((k, j)) = w_j(k)$ .

Given a word  $w$  indexed by  $J$  and two cuts  $c_1 = (K_1, L_1)$  and  $c_2 = (K_2, L_2)$  of  $J$  such that  $c_1 \leq_{\widehat{J}} c_2$ , the word  $w[c_1 : c_2]$  is obtained by restricting  $w$  to the linear ordering  $L_1 \cap K_2$ . In the case where  $c_1 = c_2$ , the word  $w[c_1 : c_2]$  is thus the empty word.

► **Example 2.** The word  $w : \mathbb{N} + \mathbb{Z} + 2 \rightarrow \{a, b\}$  that maps a single element of the part  $\mathbb{Z}$  to the letter  $b$  and the remaining elements to the letter  $a$  is illustrated in Figure 2. Dots represent the elements of  $\mathbb{N} + \mathbb{Z} + 2$ , vertical bars are its cuts, and the letters are the result of the mapping. Note that the value, in  $\mathbb{Z}$ , of the element mapped to  $b$  is irrelevant since every possible choice leads to a word isomorphic to  $w$ .



■ **Figure 2** Example of word indexed by the linear ordering  $\mathbb{N} + \mathbb{Z} + 2$ .

## 2.3 A Few Facts About Ordinals

An *ordinal* is a linear ordering  $\alpha$  such that any non-empty subordering of  $\alpha$  has a least element. In particular, this implies that each element  $j \in \alpha$  that is not the greatest element has a successor in  $\alpha$ . The class of all the ordinals is linearly ordered. Informally speaking, given two ordinals  $\alpha$  and  $\beta$ , we have  $\alpha < \beta$  if  $\alpha$  is isomorphic to a strict initial segment of  $\beta$ . That is, if there exists  $b \in \beta$  such that an order-preserving bijection exists between  $\alpha$  and  $\{j \in \beta \mid j <_{\beta} b\}$ . We refer to [21] for a thorough and formal introduction to ordinals and cardinals.

In this work, ordinals mainly appear as cofinalities of linear orderings and are denoted by Greek letters. Given a linear ordering  $J$ , a subordering  $I \subseteq J$  is *cofinal* in  $J$  if for all  $j \in J$ , there exists  $i \in I$  such that  $j \leq i$ . The *cofinality* of  $J$ , denoted by  $cf(J)$ , is defined as the least ordinal  $\xi$  such that there exists a function  $f : \xi \rightarrow J$  whose image is cofinal in  $J$ . The *coinitiality* of  $J$  is the cofinality of  $-J$ .

An ordinal is called *regular* if it equals the cofinality of some linear ordering. Alternatively, a regular ordinal can be defined as an ordinal equal to its cofinality. It follows from the definition of cofinality that a regular ordinal is necessarily the least ordinal of its cardinality, i.e., a regular ordinal is always *initial*. The first regular ordinals are 0 (0 is the cofinality of the empty linear ordering), 1 (1 is the cofinality of any linear ordering that has a greatest element),  $\omega$  (the first infinite ordinal),  $\omega_1$  (the first uncountable ordinal, i.e., the first ordinal of cardinality  $\aleph_1$ ) and  $\omega_2$  (the first ordinal of cardinality  $\aleph_2$ ).

Later, we will often need to consider concatenations of the form  $\prod_{j \in \alpha} w$  or  $\prod_{j \in -\alpha} w$  where  $\alpha$  is an ordinal and  $w$  is a word. These two concatenations are denoted concisely by  $w^\alpha$  and  $w^{-\alpha}$ , respectively. Note that this definition leads to the usual notion of  $\omega$ -power. Indeed, in the particular case where  $\alpha = \omega$ , we get  $\prod_{j \in \omega} w = w^\omega$ .

## 2.4 Automata on Linear Orderings

Automata on linear orderings, introduced in [5] and extended in [3, 4], are finite-state machines capable of reading words indexed by linear orderings. They can be seen as a broad generalization of infinite-word automata and transfinite-word automata.

► **Definition 3.** An automaton on linear orderings is a tuple  $\mathcal{A} = (Q, \Sigma, \Delta, I, F)$  where

- $Q$  is a finite set of states,
- $\Sigma$  is a finite alphabet,
- $\Delta \subseteq (Q \times \Sigma \times Q) \cup (Q \times 2^Q) \cup (2^Q \times Q)$  is a transition relation,
- $I \subseteq Q$  is a set of initial states, and
- $F \subseteq Q$  is a set of final states.

The transition relation  $\Delta$  is divided into three parts. The transitions of the form  $(q_1, a, q_2)$ , where  $q_1, q_2 \in Q$  and  $a \in \Sigma$ , correspond to the usual transitions in finite-word automata. They are called *successor transitions*. Those of the form  $(q, P)$ , with  $q \in Q$  and  $P \subseteq 2^Q$ , are called *right-limit transitions*. We often denote them by  $q \rightarrow P$ , and we say that  $P$  is the *right-limit set* of the transition. Symmetrically, the transitions of the form  $(P, q)$  are *left-limit transitions*, denoted by  $P \rightarrow q$ , and  $P$  is said to be the *left-limit set* of the transition.

Let us now define the semantics of automata on linear orderings.

► **Definition 4.** A run of  $\mathcal{A}$  reading a word  $w$  of length  $J$  is a function  $\rho : \hat{J} \rightarrow Q$  that satisfies the following conditions:

- $\rho(\hat{J}_{\min}) \in I$  and  $\rho(\hat{J}_{\max}) \in F$ .
- For every consecutive  $c_1 = (K_1, L_1), c_2 = (K_2, L_2) \in \hat{J}$ , i.e., such that  $K_2 = K_1 \cup \{j\}$  for some  $j \in J$ , one has  $(\rho(c_1), w(j), \rho(c_2)) \in \Delta$ .
- For every  $c \neq \hat{J}_{\min} \in \hat{J}$  without any predecessor, one has  $\lim_{c-} \rho \rightarrow \rho(c) \in \Delta$ , where  $\lim_{c-} \rho = \{q \in Q \mid \forall c_1 < c. \exists c_2. c_1 < c_2 < c \wedge \rho(c_2) = q\}$ .
- For every  $c \neq \hat{J}_{\max} \in \hat{J}$  without any successor, one has  $\rho(c) \rightarrow \lim_{c+} \rho \in \Delta$ , where  $\lim_{c+} \rho = \{q \in Q \mid \forall c_1 > c. \exists c_2. c < c_2 < c_1 \wedge \rho(c_2) = q\}$ .

Intuitively, a run  $\rho$  reading a word  $w$  of length  $J$  must assign an initial state to the first cut and a final state to the last cut. Moreover, if two cuts  $c_1$  and  $c_2$  are consecutive, there must exist a successor transition from  $\rho(c_1)$  to  $\rho(c_2)$ , and this successor transition must read the symbol comprised between  $c_1$  and  $c_2$ . The name *successor transition* refers to the fact that  $c_2$  is the successor of  $c_1$  in the ordering  $\hat{J}$ . On the other hand, if a cut  $c$  (different from the last one) does not have any successor, then there exist cuts in every right neighborhood of  $c$ . To leave this cut,  $\rho$  must follow a right-limit transition  $\rho(c) \rightarrow P$ . This is only valid if  $P$  is exactly equal to the *right-limit* of  $\rho$  at cut  $c$ , i.e., if  $P$  is equal to the set of states visited

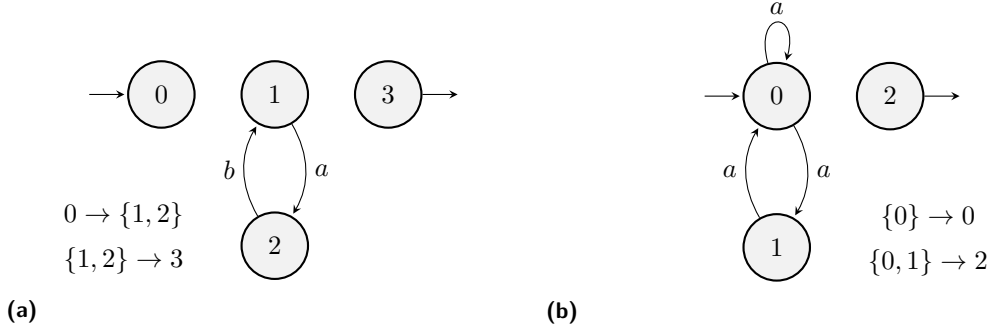
by  $\rho$  in every right neighborhood of  $c$ . The case where a cut  $c$  (different from the first cut) does not have any predecessor is symmetrical: such a cut must be reached by a left-limit transition. Note that the semantics of the limit transitions is close to the Muller acceptance condition in infinite-word automata [15], in the sense that a right-limit transition  $q \rightarrow P$  (resp. a left-limit transition  $P \rightarrow q$ ) can only be followed if  $P$  is exactly equal to the set of states visited in every right neighborhood (resp. every left neighborhood) of a cut.

An automaton on linear orderings  $\mathcal{A}$  *accepts* a word  $w$  if there exists a run reading  $w$  in  $\mathcal{A}$ . The class of all such words forms the *language accepted* by  $\mathcal{A}$ . A language is said to be *rational* if it is accepted by an automaton. In this work, we focus on rational languages of words indexed by scattered (countable and uncountable) linear orderings. An automaton accepting such a language is an *automaton on scattered linear orderings*. All the automata considered in the remainder of this article will be of this kind.

A *path* of  $\mathcal{A}$  is similar to a run, except that it is not forced to start in an initial state and to end in a final one. We say that a word  $w$  is *read* by  $\mathcal{A}$  if there exists a path reading it in  $\mathcal{A}$ . In what follows, we will often need to state that a path  $\pi$  reading a word  $w$  of length  $J$  starts in a state  $p \in Q$ , ends in a state  $q \in Q$  and *visits exactly* the set of states  $R \subseteq Q$ , i.e.,

$$p = \pi(\hat{J}_{\min}), R = \{\pi(c) \mid c \in \hat{J}\}, q = \pi(\hat{J}_{\max}).$$

We use the notation  $\pi : p \xrightarrow{R} q$  to denote this concisely.



■ **Figure 3** Examples of automata on scattered linear orderings.

► **Example 5.** A state is marked initial with an incoming arrow without origin, and final with an outgoing arrow without destination. The automaton represented in Figure 3a accepts the bi-infinite word  $(ab)^{-\omega}(ab)^{\omega}$  (since we do not distinguish isomorphic words, this is in fact the only accepted word). A run  $\rho$  reading this word can be built as follows. For any cut  $c$  immediately before a letter  $a$ , we set  $\rho(c) = 1$ . Symmetrically, for any cut  $c$  immediately before a letter  $b$ , we set  $\rho(c) = 2$ . Finally,  $\rho(\hat{J}_{\min}) = 0$  and  $\rho(\hat{J}_{\max}) = 3$ . We have  $\lim_{\hat{J}_{\min}^+} \rho = \{1, 2\} = \lim_{\hat{J}_{\max}^-} \rho$ , hence the limit transitions  $0 \rightarrow \{1, 2\}$  and  $\{1, 2\} \rightarrow 3$  can be followed to leave State 0 and reach State 3, respectively.

► **Example 6.** The automaton depicted in Figure 3b (adapted from [1]) accepts all the words  $a^\alpha$  such that  $\alpha$  is an ordinal (this follows directly from the absence of right-limit transitions) and  $\text{cf}(\alpha) = \omega$ . Since the alphabet is unary in this case, we rather say that this automaton accepts all the ordinals of cofinality  $\omega$ . Such ordinals can be accepted by using the limit transition  $\{0\} \rightarrow 0$  as many times as necessary, and by eventually following the limit transition  $\{0, 1\} \rightarrow 2$  to reach the last cut. However, if  $\alpha$  does not have the cofinality  $\omega$ , then the limit transition  $\{0, 1\} \rightarrow 2$  cannot be followed to reach State 2:

- If  $cf(\alpha) = 0$ , then  $\alpha$  is empty. Hence it is not accepted (i.e., the empty word is not accepted).
- If  $cf(\alpha) = 1$ , then  $\alpha$  has a greatest element. Therefore,  $\alpha$  cannot be accepted since State 2 does not have any incoming successor transition.
- If  $cf(\alpha) > \omega$ , the situation is less intuitive. Assume that there exists a run  $\rho$  accepting  $\alpha$  with  $cf(\alpha) = \omega_1$ . The arguments that follow also work if  $\alpha$  has another uncountable cofinality (we will prove this formally in Section 4). Since State 2 is the only final state and can only be reached by following  $\{0, 1\} \rightarrow 2$ , one must have  $\lim_{\alpha_{\max}} \rho = \{0, 1\}$ . The subordering  $\hat{\alpha}_1 \subseteq \hat{\alpha}$  of all the cuts mapped to 1 by  $\rho$  is thus cofinal in  $\hat{\alpha}^*$ . Given that  $cf(\hat{\alpha}^*) = \omega_1$ , this implies that  $cf(\hat{\alpha}_1) = \omega_1$ .

However, no transition can be taken from State 2, hence the limit transition  $\{0, 1\} \rightarrow 2$  can be followed only once. This implies that  $\hat{\alpha}_1$  is isomorphic to a limit ordinal  $\beta$  such that any  $\gamma < \beta$  is not a limit ordinal. This leads to  $\beta = \omega$ , which is a contradiction since  $cf(\omega) = \omega \neq \omega_1$ .

The key element in this reasoning is the absence of a limit transition  $\{0, 1\} \rightarrow q$  with  $q \in \{0, 1\}$ . In other words, after visiting infinitely often States 0 and 1, it is mandatory to exit  $\{0, 1\}$ . The situation would be completely different if there was the additional limit transition  $\{0, 1\} \rightarrow 0$ , for instance. In this case, all the ordinals with a cofinality at least equal to  $\omega$  would be accepted. It suggests that there are only two kinds of left-limit sets  $P$  (the same classification applies to right-limit sets): those for which there exists  $p \in P$  such that  $P \rightarrow p \in \Delta$ , and those that do not have this property and can only be used to read ordinals of cofinality  $\omega$ . The intuition is that automata on linear orderings are only able to distinguish two infinite cofinalities. We will formalize this idea in Sections 3 and 4, where we will prove that automata on scattered linear orderings are able to distinguish ordinals of cofinality greater than  $\omega$  from ones of cofinality exactly equal to  $\omega$ , but cannot discriminate two different cofinalities beyond  $\omega$ .

In fact, this example shows that automata on scattered linear orderings are not completable. Indeed, there does not exist any automaton accepting exactly the orderings that are not ordinals, together with the ordinals of cofinality different from  $\omega$ . This comes from the observation that any automaton accepting an ordinal of cofinality greater than  $\omega$  also accepts an ordinal of cofinality  $\omega$ . This property follows from [8, Theorem 2].

### 3 Constructible Words

We are now ready to define the notion of *constructible* word over a given alphabet. Intuitively, a word is constructible if it can be obtained by applying some operators to the letters of the alphabet a finite number of times. Constructible words are convenient objects to work with, since they admit a finite representation (which is not the case for unrestricted words on scattered linear orderings, since there are uncountably many such words).

► **Definition 7.** Let  $\Sigma$  be a finite alphabet. The set of constructible words  $\mathcal{C}$  on the alphabet  $\Sigma$  is the minimal set of words (in the sense of inclusion) verifying the following properties:

- The words containing a single letter  $\sigma \in \Sigma$  (i.e., the words indexed by a singleton) and the empty word  $\varepsilon$  belong to  $\mathcal{C}$ .
- If  $u, v \in \mathcal{C}$ , then  $u \cdot v \in \mathcal{C}$ .
- If  $u \in \mathcal{C}$ , then  $u^\omega \in \mathcal{C}$  and  $u^{-\omega} \in \mathcal{C}$ .
- If  $u \in \mathcal{C}$ , then  $u^{\omega_1} \in \mathcal{C}$  and  $u^{-\omega_1} \in \mathcal{C}$ .

Notice that constructible words are a generalization of *ultimately periodic words*, that are known to characterize  $\omega$ -regular languages [7]. They also generalize the constructible words introduced by Carton to study the emptiness of automata on scattered linear orderings [8,

Theorem 2]. We show in Section 5 that constructible words entirely characterize rational languages of words on scattered linear orderings. That is, two rational languages are equal if and only if they contain the same constructible words.

Observe that the only operators that are able to generate uncountable word lengths are the  $\omega_1$ -power and the reverse  $\omega_1$ -power. We will show in Section 4 that uncountable cofinalities are not distinguishable by automata on scattered linear orderings (as already mentioned in Example 6). Among all the uncountable cofinalities, we decided to choose the smallest one, i.e.,  $\omega_1$ , as representative. Symmetrically, we chose  $-\omega_1$  as representative of the reverse uncountable cofinalities.

#### 4 An Equivalence Relation on Words

We have already mentioned several times that some words cannot be distinguished by an automaton on scattered linear orderings. We now formalize this by means of an equivalence relation on words. Intuitively, given an automaton  $\mathcal{A}$ , two words are equivalent if they can be read in the exact same ways in  $\mathcal{A}$ . This idea originates from Büchi's proof that  $\omega$ -regular languages are closed under complementation [6, 23]. It was then adapted in [10] to automata on linear orderings (in the restricted case of finite-rank scattered words).

Let  $\mathcal{A} = (Q, \Sigma, \Delta, I, F)$  be an automaton on linear orderings. We say that two words  $u$  and  $v$  are *equivalent for  $\mathcal{A}$* , which we denote by  $u \sim_{\mathcal{A}} v$ , if for any  $(p, R, q) \in Q \times 2^Q \times Q$ , there exists a path  $p \xrightarrow{R} q$  reading  $u$  if and only if there exists a path  $p \xrightarrow{R} q$  reading  $v$ . When the automaton  $\mathcal{A}$  is clear from the context, we also say that  $u$  and  $v$  are *equivalent* and we write  $u \sim v$  in place of  $u \sim_{\mathcal{A}} v$ .

The relation  $\sim$  is an equivalence relation. The set of equivalence classes induced by  $\sim$  is denoted by  $S = \Sigma^{\text{scat}} / \sim$ . We denote by  $\varphi : \Sigma^{\text{scat}} \rightarrow S$  the function that maps a scattered word on the alphabet  $\Sigma$  to its equivalence class. The equivalence class  $\varphi(u)$  of a word  $u$  can be entirely described by the set of all the triples  $(p, R, q) \in Q \times 2^Q \times Q$  such that there exists a path  $p \xrightarrow{R} q$  reading  $u$ . This set of triples is denoted by  $\tilde{\varphi}(u)$ , and can be considered as an unambiguous description of the equivalence class  $\varphi(u)$ . The set of words mapped to  $s \in S$  by  $\varphi$  is denoted by  $\varphi^{-1}(s)$ .

Since  $Q$  is finite, there are only finitely many triples  $(p, R, q)$ . The number of sets of such triples is thus also finite, i.e.,  $S$  is necessarily a finite set.

► **Example 8.** Consider again the automaton in Figure 3a. In this automaton, the words  $\{(ab)^n \mid n \in \mathbb{N}_{>0}\}$  are all equivalent. For any  $n > 0$ , we have

$$\tilde{\varphi}((ab)^n) = \left\{ (1, \{1, 2\}, 1) \right\}$$

since the word  $(ab)^n$  can be read by a path  $\pi$  if and only if  $\pi$  is of the form  $1 \xrightarrow{\{1,2\}} 1$ . Now consider the words  $\{a^n \mid n \in \mathbb{N}_{>1}\}$ . Again, all these words are equivalent. In fact, they cannot be read by any path, hence  $\tilde{\varphi}(a^n) = \emptyset$  for any  $n \in \mathbb{N}_{>1}$ . Finally, notice that in this case, the language accepted by this automaton, which consists of the unique word  $(ab)^{-\omega}(ab)^{\omega}$ , is equal to  $\varphi^{-1}(s)$  where  $s$  is the equivalence class described by  $\{(0, \{0, 1, 2, 3\}, 3)\}$ .

However, given an arbitrary equivalence class  $s \in S$ , the language  $\varphi^{-1}(s)$  is not rational in general. Consider the automaton in Figure 3b. The language  $L$  of all the words  $u$  on the alphabet  $\{a\}$  such that

$$\tilde{\varphi}(u) = \left\{ (0, \{0\}, 0), (0, \{0, 1\}, 0), (1, \{0, 1\}, 0) \right\}$$

is not rational. Indeed, it contains all the words  $a^{\alpha}$  where  $\alpha$  is an ordinal such that:

- $cf(\alpha) \neq 0$  (since the empty word cannot be read by a path visiting  $\{0, 1\}$ ),
- $cf(\alpha) \neq 1$  (since any word with a last symbol can be read by a path  $\pi : 0 \xrightarrow{\{0,1\}} 1$ , but  $(0, \{0, 1\}, 1)$  does not belong to  $\tilde{\varphi}(u)$ ),
- and  $cf(\alpha) \neq \omega$  (given that any word  $a^\alpha$  with  $cf(\alpha) = \omega$  can be read by a path  $\pi : 0 \xrightarrow{\{0,1,2\}} 2$ , and  $(0, \{0, 1, 2\}, 2)$  does not belong to  $\tilde{\varphi}(u)$ ).

In fact,  $L$  is equal to the set of words  $a^\alpha$  where  $\alpha$  is an ordinal such that  $cf(\alpha) > \omega$ . As previously mentioned in Example 6, this language is not rational.

We are now ready to formalize the fact that automata on scattered linear orderings are not able to discriminate uncountable cofinalities. This is the purpose of Lemma 14. Given two regular uncountable ordinals  $\alpha$  and  $\beta$  and an automaton on scattered linear orderings  $\mathcal{A}$ , we prove that if there exists a path  $\pi$  reading a word  $w^\alpha$  in  $\mathcal{A}$ , then one can build another path  $\pi'$  that reads  $w^\beta$  and starts in the same state, ends in the same state and visits exactly the same states as  $\pi$ . Our strategy is to cut  $\pi$  into pieces (Proposition 11), to minimize them (Proposition 12) and finally to recompose them to build  $\pi'$  (Proposition 13). We start by two propositions related to ordinals. Their proofs are standard. They can be adapted to deal with reverse ordinals, by replacing the cofinality by the coinitality.

► **Proposition 9.** *Let  $\alpha$  be an ordinal with uncountable cofinality, i.e., with a cofinality at least equal to  $\omega_1$ . Then the set of limit ordinals strictly lower than  $\alpha$  is cofinal in  $\alpha$ .*

► **Proposition 10.** *If  $\alpha$  is a regular ordinal, then any subset of  $\alpha$  cofinal in  $\alpha$  is isomorphic to  $\alpha$ .*

Given a word  $w$  and an ordinal  $\alpha$ , we say that a cut of the word  $w^\alpha$  is *external* (that is, external to  $w$ ) if it separates  $w^\alpha$  into two factors  $w^{\alpha_1}$  and  $w^{\alpha_2}$  (with  $\alpha_1$  and  $\alpha_2$  two ordinal numbers, possibly equal to 0, and such that  $\alpha = \alpha_1 + \alpha_2$ ).

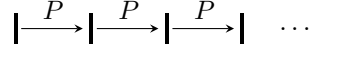
► **Proposition 11.** *Let  $\mathcal{A}$  be an automaton on linear orderings and let  $\pi$  be a path reading a word  $w^\alpha$  in  $\mathcal{A}$ , where  $w$  is a non-empty word and  $\alpha$  is an uncountable regular ordinal. Let  $P \rightarrow q$  be the limit transition followed by  $\pi$  to reach the last cut of  $w^\alpha$ . Then, there exists a state  $p$  and an ordering  $(c_i)_{i \in \alpha}$  of external cuts of  $w^\alpha$  such that:*

- $p \in P$  and  $P \rightarrow p$  is a limit transition of  $\mathcal{A}$ ,
- $\pi(c_i) = p$  for any  $i \in \alpha$ ,
- and for any  $i < j \in \alpha$ ,  $\pi$  visits  $P$  between  $c_i$  and  $c_j$ ,

**Proof.** Let  $J$  be the length of  $w^\alpha$ . The idea of the proof is to build a chain of suborderings of  $\hat{J}$  that are all cofinal in  $\hat{J}^*$  and isomorphic to  $\alpha$ . The last element of this chain will be the ordering  $(c_i)_{i \in \alpha}$ .

First, let  $C \subseteq \hat{J}^*$  be the set of non-extremal cuts external to  $w$ . The ordering  $C$  is cofinal in  $\hat{J}^*$  and isomorphic to  $\alpha$ . Next, define  $C^P \subseteq C$  as an ordering cofinal in  $C$  and such that  $\pi$  visits exactly  $P$  between any two cuts in  $C^P$ . The existence of this ordering and the fact that it is cofinal in  $C$  follow from the assumption that  $\pi$  reaches the final cut of  $w^\alpha$  thanks to the limit transition  $P \rightarrow q$ . Given that  $C^P$  is cofinal in an ordering that is isomorphic to the regular ordinal  $\alpha$ ,  $C^P$  is itself isomorphic to  $\alpha$  (thanks to Proposition 10).

Secondly, define the ordering  $C^{P,lim} \subseteq C^P$  as the set of cuts in  $C^P$  that do not have a predecessor in  $C^P$ . By Proposition 9,  $C^{P,lim}$  is cofinal in  $C^P$  (since the cofinality of  $C^P$  is  $\alpha$ , i.e., is uncountable). By Proposition 10,  $C^{P,lim}$  is thus isomorphic to  $\alpha$ . Figure 4 shows some cuts of  $C^P$ . Between any two such cuts,  $\pi$  visits exactly  $P$ . The bigger vertical bar represents a cut in  $C^{P,lim}$ .



■ **Figure 4** Several cuts of  $C^P$  and a cut of  $C^{P,lim}$ .

Each of the cuts in  $C^{P,lim}$  (except maybe the first one) is reached by  $\pi$  thanks to a left-limit transition involving  $P$  as limit set. There are finitely many such limit transitions. Hence there exists a state  $p$  and an ordering  $C^{P \rightarrow p} \subseteq C^{P,lim}$  cofinal in  $C^{P,lim}$  (and thus isomorphic to  $\alpha$  by Proposition 10) such that  $\pi$  follows the limit transition  $P \rightarrow p$  to reach each cut in  $C^{P \rightarrow p}$ . Note that we necessarily have  $p \in P$ , since the set of states visited between two cuts in  $C^{P \rightarrow p}$  is exactly  $P$ . Therefore, the ordering  $C^{P \rightarrow p}$  satisfies all the properties enumerated in the proposition. ◀

► **Proposition 12.** *Let  $\mathcal{A}$  be an automaton on scattered linear orderings with set of states  $Q$ . Let  $\alpha \neq 0$  be an ordinal number and let  $w$  be a word. Assume that there exists a path  $\pi : p \xrightarrow{R} q$  reading  $w^\alpha$  in  $\mathcal{A}$ , where  $p, q \in Q$  and  $R \in 2^Q$ . Then, there exists a path  $\pi' : p \xrightarrow{R} q$  and a countable ordinal number  $\alpha' \neq 0$  such that  $\pi'$  reads  $w^{\alpha'}$  in  $\mathcal{A}$ .*

**Proof.** We only provide a proof sketch. It can be shown that there exists an automaton  $\mathcal{B}$  with a unary alphabet  $\{\sigma\}$  that accepts the word  $\sigma^\beta$  if and only if  $w^\beta$  can be read in  $\mathcal{A}$  by a path starting in  $p$ , ending in  $q$  and visiting exactly  $R$ . By assumption, there exists a path  $\pi : p \xrightarrow{R} q$  reading the word  $w^\alpha$  in  $\mathcal{A}$ . Therefore,  $\mathcal{B}$  accepts  $\sigma^\alpha$ . The language accepted by  $\mathcal{B}$  is thus non-empty. Following the principles of the emptiness test developed in [8], one can find a word accepted by  $\mathcal{B}$  that has finite rank. This word takes the form  $\sigma^{\alpha'}$  where  $\alpha'$  is an ordinal less than  $\omega^\omega$ , hence countable. Finally, by definition of  $\mathcal{B}$ , the word  $w^{\alpha'}$  can be read in  $\mathcal{A}$  by a path  $\pi' : p \xrightarrow{R} q$ . ◀

► **Proposition 13.** *Let  $\alpha$  and  $\beta$  be two regular uncountable ordinals. Let  $w$  be a scattered word and let  $\mathcal{A}$  be an automaton on scattered linear orderings with set of states  $Q$ . If there exists a path  $\pi : q_1 \xrightarrow{R} q_2$  reading  $w^\alpha$ , then there also exists a path  $\pi' : q_1 \xrightarrow{R} q_2$  reading  $w^\beta$ .*

**Proof.** We denote by  $J$  the length of  $w$ . If  $w$  is the empty word,  $w^\alpha = w^\beta$ , hence the proposition is proven correct. We thus assume that  $w$  is not the empty word. The idea of the proof is to analyze  $\pi$ , and then to build a path  $\pi'$  reading  $w^\beta$  that keeps the properties of  $\pi$ .

First, notice that since  $\alpha$  is regular and uncountable, it is an initial ordinal greater or equal to  $\omega_1$ . Therefore it is necessarily a limit ordinal. The path  $\pi$  thus reaches the last state  $q_2$  via a limit transition  $P \rightarrow q_2$  where  $P \subseteq Q$ . Then, there exists a state  $p \in P$  and a sequence of cuts  $(c_i)_{i \in \alpha}$  with the properties enumerated in Proposition 11. Let  $\hat{J}_{min}$  be the first cut of  $w^\alpha$ . We assume without loss of generality that  $c_0 \neq \hat{J}_{min}$  (if  $c_0 = \hat{J}_{min}$ , then redefine the sequence  $(c_i)_{i \in \alpha}$  by setting  $c_0$  to  $c_1$ ,  $c_1$  to  $c_2$ , etc.). Let us define  $\pi_1$  as the restriction of  $\pi$  to the prefix  $w[\hat{J}_{min} : c_0]$ , and  $\pi_2$  as the restriction of  $\pi$  to the subword  $w[c_0 : c_1]$ . From Proposition 11, we infer the following properties of  $\pi_1$  and  $\pi_2$  (together with the fact that  $P \rightarrow p$  is a limit transition of  $\mathcal{A}$ ):

- $\pi_1$  reads  $w^{\theta_1}$  for some ordinal number  $\theta_1 > 0$ , starts in  $q_1$ , ends in  $p$  and visits exactly some set of states  $U \subseteq Q$ ,
- $\pi_2$  reads  $w^{\theta_2}$  for some ordinal number  $\theta_2 > 0$ , starts in  $p$ , ends in  $p$  and visits exactly  $P$ ,
- and  $U \cup P \cup \{q_2\} = R$ .

Now let us apply Proposition 12 twice (once for  $\pi_1$  and once for  $\pi_2$ ). This yields a path  $\pi'_1 : q_1 \xrightarrow{U} p$  reading a word  $w^{\theta'_1}$  and a path  $\pi'_2 : p \xrightarrow{P} p$  reading a word  $w^{\theta'_2}$ , where  $\theta'_1, \theta'_2$  are countable ordinal different from 0.

We are now ready to build a path  $\pi' : q_1 \xrightarrow{R} q_2$  reading  $w^\beta$ . First, notice that  $\beta > \theta'_1 > 0$  and  $\beta > \theta'_2 > 0$  (since  $\beta$  is uncountable and  $\theta'_1, \theta'_2$  are countable), and  $\beta$  is an initial ordinal (since it is regular). This implies that  $\theta'_2 \cdot \beta = \beta$  and  $\theta'_1 + \beta = \beta$  (see e.g. [19, Exercise 3.39]). In terms of words, this means that

$$w^\beta = w^{\theta'_1} \cdot (w^{\theta'_2})^\beta.$$

Now consider the path  $\pi'$  that first follows  $\pi'_1$  to read  $w^{\theta'_1}$ , then follows  $\beta$  times  $\pi'_2$  to read each copy of  $w^{\theta'_2}$ , and eventually reaches the last cut of  $w^\beta$  via the limit transition  $P \rightarrow q_2$ . This path reads  $w^\beta$ , starts in  $q_1$ , visits  $R$  and ends in  $q_2$ . Moreover, it can be checked that this path is valid. Any cut of the word  $(w^{\theta'_2})^\beta$  that is external to  $w^{\theta'_2}$  and that does not have a predecessor that is external can be reached thanks to the limit transition  $P \rightarrow p$  (or  $P \rightarrow q_2$  for the last cut). ◀

► **Lemma 14.** *Let  $\alpha$  and  $\beta$  be two regular uncountable ordinals. For any automaton on scattered linear orderings  $\mathcal{A}$  and for any scattered word  $w$ , the words  $w^\alpha$  and  $w^\beta$  are equivalent for  $\mathcal{A}$ . Symmetrically,  $w^{-\alpha}$  and  $w^{-\beta}$  are also equivalent for  $\mathcal{A}$ .*

**Proof.** It follows from the fact that Proposition 13 is symmetric regarding  $\alpha$  and  $\beta$ . Hence, the set of triples  $(q_1, R, q_2) \in Q \times 2^Q \times Q$  that characterize the equivalence class of  $w^\alpha$  are the same as those characterizing the equivalence class of  $w^\beta$ , i.e.,  $w^\alpha$  and  $w^\beta$  are in the same equivalence class.

The proof that  $w^{-\alpha}$  and  $w^{-\beta}$  are equivalent for  $\mathcal{A}$  is essentially the same. It relies on the mirrored versions of Propositions 9 to 13. ◀

The next step is to understand how the equivalence classes induced by  $\sim_{\mathcal{A}}$  behave with respect to the concatenation of words. Lemma 15 addresses this question. It will be used intensively in Section 5.

► **Lemma 15.** *Let  $\mathcal{A}$  be an automaton on scattered linear orderings and let  $I$  be a non-empty scattered linear ordering. Let  $\{u_i \mid i \in I\}$  and  $\{v_i \mid i \in I\}$  be two sets of words such that  $u_i \sim_{\mathcal{A}} v_i$  for all  $i \in I$ . Then the words  $u$  and  $v$  respectively defined as  $(\prod_{i \in I} u_i)$  and  $(\prod_{i \in I} v_i)$  are equivalent for  $\mathcal{A}$ .*

**Proof.** We only give a proof sketch. Given two paths  $\pi$  and  $\tau$  in  $\mathcal{A}$ , we write  $\pi \equiv \tau$  if both paths start in the same state, end in the same state, and visit exactly the same set of states. Since  $u_i \sim_{\mathcal{A}} v_i$ , the existence of a path  $\pi_i$  reading  $u_i$  in  $\mathcal{A}$  implies that there also exists a path  $\tau_i \equiv \pi_i$  reading  $v_i$  in  $\mathcal{A}$ . The idea of the proof is thus to consider a path  $\pi$  reading  $u = \prod_{i \in I} u_i$  in  $\mathcal{A}$  and then to split it into a set of paths  $\{\pi_i \mid i \in I\}$  such that  $\pi_i$  reads  $u_i$ . From this, we can infer the existence of a set of paths  $\{\tau_i \mid i \in I\}$  such that  $\tau_i$  reads  $v_i$  and  $\tau_i \equiv \pi_i$ , and finally build a path  $\tau \equiv \pi$  reading  $v$  in  $\mathcal{A}$ . By symmetry of the problem, this is sufficient to show that  $u$  and  $v$  are equivalent for  $\mathcal{A}$ . ◀

In the particular case where  $\alpha = 2$ , Lemma 15 actually shows that it makes sense to define a binary product operation between the equivalence classes induced by  $\sim_{\mathcal{A}}$ . Given two equivalence classes  $s_1$  and  $s_2$ , their product  $s_1 \cdot s_2$  is defined as the equivalence class of any word  $u_1 \cdot u_2$  such that  $u_1$  belongs to  $s_1$  and  $u_2$  to  $s_2$ . Equivalently,

$$\varphi(u_1) \cdot \varphi(u_2) = \varphi(u_1 \cdot u_2). \quad (1)$$

for any  $u_1, u_2 \in \Sigma^{\text{scat}}$ . Notice that the product we defined on  $S$  is an associative operation, since

$$\varphi(u_1) \cdot (\varphi(u_2) \cdot \varphi(u_3)) = \varphi(u_1 \cdot u_2 \cdot u_3) = (\varphi(u_1) \cdot \varphi(u_2)) \cdot \varphi(u_3).$$

Therefore, the set  $S$  containing the equivalence classes induced by  $\sim_{\mathcal{A}}$  has a finite-semigroup structure, since it is a finite set equipped with a binary product operation that is associative. Moreover, it follows from Equation 1 that  $\varphi : \Sigma^{\text{scat}} \rightarrow S$  is a *semigroup morphism* from  $\Sigma^{\text{scat}}$  (which can be seen as an infinite semigroup where the associative operation is the concatenation) to the finite semigroup  $S$ .

The links between automata and semigroups are well studied in the literature. In [18], rational languages of words indexed by countable scattered linear orderings are shown to be exactly the languages *recognizable* by a special form of finite semigroups. That is, a language  $L \subseteq \Sigma^\diamond$  (where  $\Sigma^\diamond$  is the class of all the words on the alphabet  $\Sigma$  indexed by countable scattered linear orderings) is rational if and only if it can be written as  $\bigcup_{t \in T} \varphi^{-1}(t)$  where  $T$  is a subset of a finite semigroup  $S$  and  $\varphi : \Sigma^\diamond \rightarrow S$  is a semigroup morphism (i.e.,  $\varphi$  satisfies Equation 1). This result generalizes the correspondence between regular languages of finite words and finite semigroups [13]. It also proves that rational languages of words indexed by countable scattered linear orderings are closed under complementation.

These results do not extend to rational languages of scattered words (countable and uncountable). Given a finite semigroup  $S$  and a semigroup morphism  $\varphi : \Sigma^{\text{scat}} \rightarrow S$ , the languages  $\varphi^{-1}(s)$  (where  $s \in S$ ) are not rational in general, hence the equivalence between rationality and recognizability by semigroups does not hold in our settings. If it were the case, then automata on scattered linear orderings would be complementable. Nevertheless, it is still possible to exploit the fact that the set of equivalence classes induced by  $\sim_{\mathcal{A}}$  has a finite-semigroup structure. This is the objective of Section 5.

## 5 Constructible Words Characterize Rational Languages of Scattered Words

In this section, we prove that the language accepted by an automaton on scattered linear orderings  $\mathcal{A}$  is characterized by the constructible words accepted by  $\mathcal{A}$  (Theorem 20). We first prove Theorem 19, stating that for any non-empty equivalence class  $s$  induced by  $\sim_{\mathcal{A}}$ , the language  $\varphi^{-1}(s)$  contains a constructible word (despite being non-rational in general). The proof of this result relies on the application of a theorem of Colcombet [12]. We introduce this theorem in Section 5.1 and apply it to our problem in Section 5.2.

► **Remark 16.** In the restricted case of automata on countable scattered linear orderings, Theorem 19 follows immediately from the fact that each equivalence class is a rational language [18, Proposition 27], and any non-empty rational language contains a countable constructible word [8, Theorem 2]. Alternatively, the fact that the semigroups introduced in [18] (as an algebraic model equivalent to automata on countable scattered linear orderings) can be described by an algebra whose operators are the concatenation, the  $\omega$ -power and the reverse  $\omega$ -power, yields another proof of Theorem 19 in the restricted case of countable scattered words. In this restricted case, automata can be complemented [18], hence Theorem 20 can also be directly proven by adapting the proof of the characterization of the  $\omega$ -regular languages by their ultimately periodic words [7, Fact 1].

### 5.1 A Ramsey-like Theorem for Semigroups

The theorem presented in this section comes from [12]. It can be seen as an extension of Ramsey's theorem [17] to graphs whose edges are colored according to a semigroup structure. Its proof relies on the axiom of choice.

Let  $J$  be a linear ordering and let  $S$  be a finite semigroup. The product of two elements  $s_1, s_2 \in S$  is denoted by  $s_1 \cdot s_2$ . A *multiplicative labeling*  $\sigma$  is a function that maps each pair  $(i, j) \in J \times J$  such that  $i < j$  to an element of  $S$ , and such that

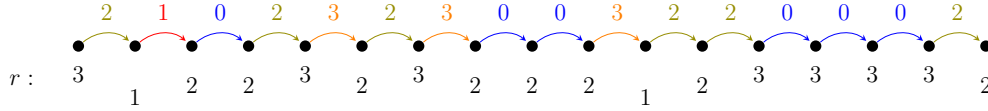
$$\sigma(i, j) \cdot \sigma(j, k) = \sigma(i, k)$$

for all  $i < j < k \in J$ . In other words, if we consider  $J$  as an ordered set of vertices of a complete graph, the multiplicative labeling  $\sigma$  is a coloring of the edges of this graph, where the colors satisfy a semigroup structure (which is not the case in the classical infinite form of Ramsey's theorem [17]).

A *split* of  $J$  is a function  $r : J \rightarrow \{1, 2, \dots, N\}$  where  $N \in \mathbb{N}$  is called the *height* of the split. Given a split  $r$  and an integer  $m \in \{1, 2, \dots, N\}$ , two elements  $i, j \in J$  are said to be *m-neighbors* if  $r(i) = r(j) = m$  and  $r(k) \leq m$  for all  $k$  between  $i$  and  $j$ . The relation “being *m-neighbors*” is in fact an equivalence relation, for any  $m \in \{1, 2, \dots, N\}$ . An *m-class* is an equivalence class induced by this equivalence relation. A split  $r : J \rightarrow \{1, 2, \dots, N\}$  is *Ramseyan* for the multiplicative labeling  $\sigma$  if for any  $m \in \{1, 2, \dots, N\}$  and for any *m-class*  $X$ , there exists an idempotent element  $e \in S$  such that  $\sigma(i, j) = e$  for all  $i < j \in X$  (an element  $e \in S$  is *idempotent* if  $e \cdot e = e$ ).

► **Theorem 17** (Colcombet [12]). *For any linear ordering  $J$ , for any finite semigroup  $S$  and for any multiplicative labeling  $\sigma : J \times J \rightarrow S$ , there exists a split of  $J$  of height at most  $2|S|$  that is Ramseyan for  $\sigma$ .*

► **Example 18.** The following example comes from [12]. Consider the graph with  $J = 17$  vertices represented in Figure 5. The edges are labeled by the elements of the semigroup  $S = \mathbb{Z}/5\mathbb{Z}$  (the set of integers modulo 5, equipped with addition modulo 5, that we denote



■ **Figure 5** Example of Ramseyan split.

by  $+$  here). This semigroup is in fact a group, but it does not matter in this example. Note that all the edges are not represented, but the label of any edge can be retrieved thanks to the multiplicative property of the labeling. For instance, the edge linking the first vertex to the third one is labeled by  $2 + 1 = 3$ .

Theorem 17 predicts the existence of a Ramseyan split of height at most  $2|S| = 10$ . In fact, in this case there exists a Ramseyan split  $r$  of height 3, represented below the graph in Figure 5. Indeed, consider a 3-class (there is only one such class, which is the set of vertices mapped to 3). Two distinct vertices in this class are always linked by an edge labeled by the idempotent element 0. Now consider a 2-class (this time, there exist 4 such classes, two of them being a singleton). Again, two distinct elements in a 2-class are linked by an edge labeled by 0 (in fact, 0 is the only idempotent element of the semigroup). Finally, there are two 1-classes that are both singletons. Hence one cannot find two distinct vertices in these 1-classes, which implies that the Ramseyan condition is automatically satisfied.

## 5.2 Application to Constructible Words

Let  $\mathcal{A} = (Q, \Sigma, \Delta, I, F)$  be an automaton on linear orderings and let  $S$  be the set of equivalence classes induced by  $\sim_{\mathcal{A}}$ . We regard  $S$  as a finite semigroup, as established in Section 4. We denote by  $\varphi : \Sigma^{\text{scat}} \rightarrow S$  the semigroup morphism mapping a scattered word to its equivalence class.

Let  $w$  be a word indexed by a linear ordering  $J$ . We now build a multiplicative labeling of  $\widehat{J}$  from the equivalence classes of each subword of  $w$  (this idea is already present in [12]). More precisely, we define the multiplicative labeling  $\sigma : \widehat{J} \times \widehat{J} \rightarrow S$  as follows:

$$\sigma(c_1, c_2) = \varphi(w[c_1 : c_2])$$

for any  $c_1 < c_2 \in \widehat{J}$ . Thanks to Theorem 17, we know the existence of a split  $r : \widehat{J} \rightarrow \{1, 2, \dots, N\}$  whose height  $N$  is at most  $2|S|$  and that is Ramseyan for  $\sigma$ . We refer to this Ramseyan split as a *Ramseyan split* of  $w$ .

► **Theorem 19.** *Let  $\mathcal{A} = (Q, \Sigma, \Delta, I, F)$  be an automaton on scattered linear orderings. For any non-empty equivalence class  $s$  induced by  $\sim_{\mathcal{A}}$ , the language  $\varphi^{-1}(s)$  contains a constructible word.*

**Proof.** The main idea is to consider an arbitrary word  $w \in \Sigma^{\text{scat}}$  of length  $J$  and to show that  $w$  is equivalent to a constructible word. This shows that the equivalence class of  $w$  contains at least one constructible word.

Given a Ramseyan split  $r$  of  $w$ , we proceed by induction on the *internal height* of  $r$ , defined as

$$N^* = \begin{cases} \max\{r(c) \mid c \in \widehat{J}^*\} & \text{if } \widehat{J}^* \text{ is non-empty} \\ 0 & \text{otherwise.} \end{cases}$$

Since the height of  $r$  is finite by Theorem 17, so is its internal height. In what follows, we say that  $w$  is of internal height  $N^*$  if it admits a Ramseyan split of internal height  $N^*$ .

Let  $w$  be a word of internal height 0. The set of non-extremal cuts  $\widehat{J}^*$  of  $w$  is empty, hence  $w$  contains 0 or 1 letter. Hence in both cases,  $w$  is constructible. In particular,  $w$  is thus equivalent to a constructible word.

Now, let  $w$  be a word of internal height  $N^* > 0$  and  $r$  a Ramseyan split of  $w$  with internal height  $N^*$ . Let  $C \subseteq \widehat{J}$  be the set of cuts of  $w$  mapped to  $N^*$  by  $r$ . Formally,

$$C = \{c \in \widehat{J} \mid r(c) = N^*\}.$$

Importantly, all the elements of  $C$  are in the same  $N^*$ -class (indeed, they are all mapped to  $N^*$  by  $r$ , and by maximality of  $N^*$ , any element of  $\widehat{J}$  between two elements of  $C$  is mapped to  $k \leq N^*$ ). Note that  $C$  is necessarily non-empty, since it contains at least one non-extremal cut (otherwise,  $r$  would not be a split of internal height  $N^* > 0$ ). We now distinguish two cases, depending on whether  $C$  is a singleton or not. In each case, the idea of the proof is to use some cuts belonging to  $C$  as separators to factorize  $w$ .

■ If  $C$  contains a single element  $c$ , consider the following factorization of  $w$ :

$$w = w[\widehat{J}_{\min} : c] \cdot w[c : \widehat{J}_{\max}].$$

The words  $w[\widehat{J}_{\min} : c]$  and  $w[c : \widehat{J}_{\max}]$  are of internal height at most  $N^* - 1$  (since none of the non-extremal cuts of  $w[\widehat{J}_{\min} : c]$  and  $w[c : \widehat{J}_{\max}]$  is mapped to  $N^*$ ). Hence by induction hypothesis, they admit equivalent constructible words  $u_1$  and  $u_2$ , respectively. Therefore, by Lemma 15, the word  $u_1 \cdot u_2$  is equivalent to  $w$ , and it is a constructible word.

■ If  $C$  contains strictly more than one element, then it contains two elements that are consecutive in  $C$ . Let  $c < d$  be these two elements. Their existence relies on the fact that  $J$  is scattered, hence  $\widehat{J}$  is also scattered, and any subordering of  $\widehat{J}$  as well. From the

existence of  $c$  and  $d$ , one can infer the existence of a word  $w[c : d]$  of internal height at most  $N^* - 1$ . By induction hypothesis, this word thus admits an equivalent constructible word, that we denote by  $v$ . It remains to show how to use the existence of  $v$  to build a constructible word equivalent to  $w$ . We need to distinguish four cases, depending on the presence or the absence of a greatest/least element in  $C$ .

- Case 1:  $C$  has a greatest and a least element, denoted respectively by  $c_{max}$  and  $c_{min}$ . Then, consider the following factorization of  $w$ :

$$w = w[\hat{J}_{min} : c_{min}] \cdot w[c_{min} : c_{max}] \cdot w[c_{max} : \hat{J}_{max}].$$

Since  $c_{min}, c_{max}, c$  and  $d$  belong to the same  $N^*$ -class,  $\sigma(c_{min}, c_{max}) = \sigma(c, d)$  by definition of a Ramseyan split, hence the word  $w[c_{min} : c_{max}]$  is equivalent to  $w[c : d]$ , which is itself equivalent to the constructible word  $v$ . Moreover, the words  $w[\hat{J}_{min} : c_{min}]$  and  $w[c_{max} : \hat{J}_{max}]$  are of internal height at most  $N^* - 1$ , since none of their non-extremal cuts is mapped to  $N^*$  by the split  $r$ . Hence by induction hypothesis, there exist two constructible words  $u_1$  and  $u_2$  respectively equivalent to  $w[\hat{J}_{min} : c_{min}]$  and  $w[c_{max} : \hat{J}_{max}]$ . Therefore, thanks to Lemma 15,  $w$  is equivalent to  $u_1 \cdot v \cdot u_2$ , which is constructible.

- Case 2:  $C$  has a least element  $c_{min}$  but does not have any greatest element. Let  $\xi$  be the cofinality of  $C$ , and  $c_{sup} \in \hat{J}$  its least upper bound (which necessarily exists since  $\hat{J}$  is *complete*, as stated in [5]). Consider a subordering  $(c_i)_{i \in \xi}$  of  $C$  that is cofinal in  $C$ , and such that  $c_0 = c_{min}$ . Now factorize  $w$  as follows:

$$w = w[\hat{J}_{min} : c_{min}] \cdot \left( \prod_{i \in \xi} w[c_i : c_{i+1}] \right) \cdot w[c_{sup} : \hat{J}_{max}].$$

The situation is similar to the previous case. The words  $w[\hat{J}_{min} : c_{min}]$  and  $w[c_{sup} : \hat{J}_{max}]$  are of internal height at most  $N^* - 1$ , hence by induction hypothesis they admit equivalent constructible words, say  $u_1$  and  $u_2$ . Moreover,  $\sigma(c_i, c_{i+1}) = \sigma(c, d)$  for all  $i \in \xi$  since  $c, d, c_i$  and  $c_{i+1}$  are in the same  $N^*$ -class. Subsequently, each subword  $w[c_i : c_{i+1}]$  is equivalent to  $w[c, d]$ , which is itself equivalent to  $v$ . Therefore, by Lemma 15,  $w$  is equivalent to  $u_1 \cdot v^\xi \cdot u_2$ , where  $u_1, v$  and  $u_2$  are all constructible.

Note that  $\xi$  is larger or equal to the ordinal  $\omega$  (if it was not the case,  $C$  would be empty or would have a greatest element). If  $\xi = \omega$ , then  $u_1 \cdot v^\xi \cdot u_2 = u_1 \cdot v^\omega \cdot u_2$  is constructible, since the  $\omega$ -power is a constructible operator. If  $\xi > \omega$ , then  $\xi$  is an uncountable regular ordinal. Lemma 14 then ensures that  $u_1 \cdot v^\xi \cdot u_2$  is equivalent to  $u_1 \cdot v^{\omega_1} \cdot u_2$ , which is constructible. In both cases,  $w$  is equivalent to a constructible word.

- Case 3:  $C$  has a greatest element  $c_{max}$  but does not have any least element. This case is the mirrored version of the previous one. With the same arguments as before, it is possible to prove that  $w$  is equivalent to  $u_1 \cdot v^{-\omega} \cdot u_2$  or  $u_1 \cdot v^{-\omega_1} \cdot u_2$ , where  $u_1, u_2$  and  $v$  are constructible.
- Case 4:  $C$  does not have a greatest element nor a least element. This case is the combination of the two previous ones. Following the same reasoning as above, it can be proven that  $w$  is equivalent to  $u_1 \cdot v^{-\omega} \cdot v^\omega \cdot u_2$ , to  $u_1 \cdot v^{-\omega} \cdot v^{\omega_1} \cdot u_2$ , to  $u_1 \cdot v^{-\omega_1} \cdot v^\omega \cdot u_2$  or to  $u_1 \cdot v^{-\omega_1} \cdot v^{\omega_1} \cdot u_2$ , where  $u_1, u_2$  and  $v$  are constructible.

This concludes the proof. ◀

► **Theorem 20.** *Let  $\mathcal{A}_1$  and  $\mathcal{A}_2$  be two automata on scattered linear orderings accepting the same constructible words. Then,  $\mathcal{A}_1$  and  $\mathcal{A}_2$  accept the same language.*

**Proof.** Let  $\mathcal{A}$  be the automaton obtained by mere juxtaposition of  $\mathcal{A}_1$  and  $\mathcal{A}_2$  (i.e.,  $\mathcal{A}$  is the union of  $\mathcal{A}_1$  and  $\mathcal{A}_2$ ). We assume that the states of  $\mathcal{A}_1$  and  $\mathcal{A}_2$  are disjoint. If two words  $u$  and  $v$  are equivalent for  $\mathcal{A}$ , then these two words are also equivalent for  $\mathcal{A}_1$  and for  $\mathcal{A}_2$ . Let  $w$  be a word accepted by  $\mathcal{A}_1$ . Our goal is to show that  $w$  is also accepted by  $\mathcal{A}_2$ . Thanks to Theorem 19, there exists a constructible word  $u$  such that  $u \sim_{\mathcal{A}} w$ . This implies that  $u \sim_{\mathcal{A}_1} w$  and  $u \sim_{\mathcal{A}_2} w$ . Therefore,  $u$  is also accepted by  $\mathcal{A}_1$ . Since  $u$  is constructible, it is thus accepted by  $\mathcal{A}_2$  as well. Finally, since  $u \sim_{\mathcal{A}_2} w$ , the word  $w$  is also accepted by  $\mathcal{A}_2$ . ◀

► **Remark 21.** In the proof of Theorem 19, we actually showed that any word  $w$  admits an equivalent constructible word whose notation involves a bounded number of constructible operators. This comes from the fact that Colcombet’s theorem guarantees the existence of a Ramseyan split of  $w$  that is not only of finite height, but also of bounded height. Therefore, Theorem 20 could actually be slightly strengthened. To show that two automata on scattered linear orderings accept the same language, it is not necessary to check that they accept the same constructible words. It suffices to show that it is the case for constructible words involving a bounded number of operators. We believe that this slight improvement might be important in future applications of this theorem.

## 6 Conclusion and Further Work

We proved that rational languages of words indexed by scattered linear orderings are characterized by their constructible words. This generalizes the characterization of  $\omega$ -regular languages by their ultimately periodic words. The proof of our result is interesting in itself, since it does not rely on complementation but rather exploits semigroup theory.

We expect our result to be used in future developments. It is likely to simplify some proofs, since constructible words are simple objects to work with, in the sense that they admit a finite notation (which is not the case of non-constructible words). We also believe that it sheds light on the expressiveness of automata on scattered linear orderings.

It would be interesting to understand the connections between our result and other theorems of the same flavor in logic. For instance, a result of Lauchli and Leonard states that two formulas in the monadic first-order theory of order (MFO) are equivalent on scattered domains (i.e., have the same scattered models) if and only if they have the same simple (“constructible”) countable scattered models [14] (see also [19, Corollary 7.10]). To the best of our knowledge, it is not known whether a language definable in MFO on unrestricted scattered domains is always rational. Hence it is not clear whether this result of Lauchli and Leonard aligns with Theorem 20, or is of a different nature. We would like to answer this question in some future work, by studying the languages definable in MFO.

Our result could potentially be extended to rational languages over unrestricted linear orderings. It would also be worthwhile to investigate whether it yields decidability results (for instance, regarding equivalence checking between two automata on scattered linear orderings).

---

## References

- 1 Nicolas Bedon. Finite automata and ordinals. *Theoretical Computer Science*, 156(1):119–144, 1996. doi:10.1016/0304-3975(95)00006-2.

- 2 Nicolas Bedon, Alexis Bès, Olivier Carton, and Chloe Rispal. Logic and rational languages of words indexed by linear orderings. *Theory Comput. Syst.*, 46(4):737–760, 2010. doi:10.1007/s00224-009-9222-6.
- 3 Alexis Bès and Olivier Carton. A Kleene theorem for languages of words indexed by linear orderings. *International Journal of Foundations of Computer Science*, 17(03):519–541, 2006. doi:10.1142/S0129054106003978.
- 4 Bernard Boigelot, Thomas Braipson, and Tom Clara. Epsilon automata on linear orderings. In Giuseppa Castiglione and Sabrina Mantaci, editors, *Implementation and Application of Automata - 29th International Conference, CIAA 2025, Palermo, Italy, September 22-25, 2025, Proceedings*, volume 15981 of *Lecture Notes in Computer Science*, pages 41–54, Cham, 2025. Springer. doi:10.1007/978-3-032-02602-6\_4.
- 5 Véronique Bruyère and Olivier Carton. Automata on linear orderings. *Journal of Computer and System Sciences*, 73(1):1–24, 2007. doi:10.1016/j.jcss.2006.10.009.
- 6 Julius R. Büchi. On a decision method in restricted second order arithmetic. In *Proc. International Congress on Logic, Methodology and Philosophy of Science*, pages 1–12, Stanford, 1962. Stanford University Press.
- 7 Hugues Calbrix, Maurice Nivat, and Andreas Podelski. Ultimately periodic words of rational  $w$ -languages. In Stephen D. Brookes, Michael G. Main, Austin Melton, Michael W. Mislove, and David A. Schmidt, editors, *Mathematical Foundations of Programming Semantics, 9th International Conference, New Orleans, LA, USA, April 7-10, 1993, Proceedings*, volume 802 of *Lecture Notes in Computer Science*, pages 554–566, Berlin, Heidelberg, 1993. Springer. doi:10.1007/3-540-58027-1\_27.
- 8 Olivier Carton. Accessibility in automata on scattered linear orderings. In *Mathematical Foundations of Computer Science 2002*, pages 155–164. Springer Berlin Heidelberg, 2002. doi:10.1007/3-540-45687-2\_12.
- 9 Olivier Carton, Thomas Colcombet, and Gabriele Puppis. An algebraic approach to MSO-definability on countable linear orderings. *The Journal of Symbolic Logic*, 83(3):1147–1189, 2018. doi:10.1017/jsl.2018.7.
- 10 Olivier Carton and Chloé Rispal. Complementation of Rational sets on Scattered Linear Orderings of Finite Rank. *Theoretical Computer Science*, 382(2):109–119, 2007. doi:10.1016/j.tcs.2007.03.008.
- 11 Yaacov Choueka. Finite automata, definable sets, and regular expressions over  $\omega^n$ -tapes. *J. Comput. Syst. Sci.*, 17(1):81–97, 1978. doi:10.1016/0022-0000(78)90036-3.
- 12 Thomas Colcombet. Factorization forests for infinite words and applications to countable scattered linear orderings. *Theoretical Computer Science*, 411(4):751–764, 2010. Fundamentals of Computation Theory. doi:10.1016/j.tcs.2009.10.013.
- 13 Samuel Eilenberg. *Automata, languages, and machines., B*, volume B of *Pure and applied mathematics*. Academic Press, New York, 1976. URL: <https://www.worldcat.org/oclc/310535259>.
- 14 Hans Läuchli and John Leonard. On the elementary theory of linear order. *Fundamenta Mathematicae*, 59(1):109–116, 1966.
- 15 David E. Muller. Infinite sequences and finite machines. In *4th Annual Symposium on Switching Circuit Theory and Logical Design, Chicago, Illinois, USA, October 28-30, 1963*, SWCT '63, pages 3–16, USA, 1963. IEEE Computer Society. doi:10.1109/SWCT.1963.8.
- 16 Maurice Nivat and Dominique Perrin. Ensembles reconnaissables de mots biinfinis. In *Proc. 14th Annual ACM Symposium on Theory of Computing (STOC)*, pages 47–59. ACM, 1982. doi:10.1145/800070.802176.
- 17 Frank P. Ramsey. On a problem of formal logic. In *Proc. of the London Mathematical Society*, volume s2-30, pages 264–286, 1930.
- 18 Chloé Rispal and Olivier Carton. Complementation of rational sets on countable scattered linear orderings. *International Journal of Foundations of Computer Science*, 16(04):767–786, 2005. doi:10.1142/S0129054105003285.

- 19 Joseph G. Rosenstein. *Linear orderings*. Academic press, 1982.
- 20 Saharon Shelah. The monadic theory of order. *The Annals of Mathematics*, 102(3):379, 1975.
- 21 Wacław Sierpiński. *Cardinal and ordinal numbers*. Monografie matematyczne / Polska Akademia Nauk ; tom 34. Państwowe Wydawnictwo Naukowe, Warszawa, 2nd ed. rev. edition, 1965.
- 22 Imre Simon. Factorization forests of finite height. *Theoretical Computer Science*, 72(1):65–94, 1990. doi:10.1016/0304-3975(90)90047-L.
- 23 Wolfgang Thomas. Chapter 4 - automata on infinite objects. In Jan Van Leeuwen, editor, *Formal Models and Semantics*, Handbook of Theoretical Computer Science, pages 133–191. Elsevier, Amsterdam, 1990. doi:10.1016/b978-0-444-88074-1.50009-3.
- 24 Jerzy Wojciechowski. Finite automata on transfinite sequences and regular expressions. *Fundamenta Informaticae*, 8(3–4):379–396, 1985.