Factorisation forests for infinite words
Application to countable scattered linear orderings

Thomas Colcombet

CNRS/IRISA
thomas.colcombet@irisa.fr

Abstract. The theorem of *factorisation forests* shows the existence of nested factorisations — a la Ramsey — for finite words. This theorem has important applications in semigroup theory, and beyond.
We provide two improvements to the standard result. First we improve on all previously known bounds for the standard theorem. Second, we extend it to every 'complete linear ordering'. We use this variant in a simplified proof of complementation of automata over words of countable scattered domain.

Keywords: Formal languages, semigroups, infinite words, automata.

1 Introduction

Factorisation forests were introduced by Simon [15]. The associated theorem — which we call the theorem of factorisation forests below — states that for every semigroup morphism from words to a finite semigroup S, every word has a Ramseyan factorisation tree of height linearly bounded by $|S|$ (see below). An alternative presentation states that for every morphism φ from A^+ to some finite semigroup S, there exists a regular expression evaluating to A^+ in which the Kleene star L^* is allowed only when $\varphi(L) = \{e\}$ for some $e = e^2 \in S$; i.e. the Kleene star is allowed only if it produces a Ramseyan factorisation of the word.

The theorem of factorisation forests provides a very deep insight on the structure of finite semigroups, and has therefore many applications. Let us cite some of them. Distance automata are nondeterministic finite automata mapping words to naturals. An important question concerning them is the limitedness problem: decide whether this mapping is bounded or not. It has been shown decidable by Simon using the theorem of factorisation forests [15]. This theorem also allows a constructive proof of Brown's lemma on locally finite semigroups [2]. It is also used in the characterisation of subfamilies of the regular languages, for instance the polynomial closure of varieties in [11]. Or to give general characterisations of finite semigroups [10]. In the context of languages of infinite words indexed by ω, it has also been used in a complementation procedure [1] extending Buchi's lemma [4]. In [7], a deterministic variant of the theorem of factorisation forest is used for proving that every monadic second-order interpretation is equivalent over trees to the composition of a first-order interpretation and a monadic

second-order marking. This itself provides new result in the theory of finitely presentable infinite structures.

The present paper aims first at advertising the theorem of factorisation forest which, though already used in many papers, is in fact known only to a quite limited community. The reason for this is that its proofs rely on the use of Green's relations: Green's relations form an important tool in semigroup theory, but are technical and uncomfortable to work with. The merit of the factorisation forest theorem is that it is usable without any significant knowledge of semigroup theory, while it encapsulates nontrivial parts of this theory. Furthermore, as briefly mentioned above, this theorem has natural applications in automata theory.

This paper contains three contributions. First, we provide a new proof of the original theorem improving on all previously known bounds in [15] and [6]. Second, we extend the result to the infinite case (i.e., to infinite words, though we use a different presentation). Third, we use this last extension in a simplified proof of complementation of automata on countable scattered linear orderings, a result known from Carton and Rispal [5].

The content of the paper is organised as follows. Section 2 is dedicated to definitions. Section 3 presents the original theorem of factorisation forests as well as a variant in terms of Ramseyan splits and its extension to the infinite case. In Section 4 we apply this last extension to the complementation of automata over countable scattered linear orderings.

2 Definitions

In this section, we successively present linear orderings, words indexed by them, semigroups and additive labellings.

2.1 Linear orderings

A *linear ordering* $\alpha = (L, <)$ is a set L equipped with a total ordering relation $<$; i.e., an irreflexive, antisymmetric and transitive relation such that for every distinct elements x, y in L, either $x < y$ or $y < x$. Two linear orderings $\alpha = (L, <)$ and $\beta = (L', <')$ have same *order type* if there exists a bijection f from L onto L' such that for every x, y in L, $x < y$ iff $f(x) <' f(y)$. We denote by $\omega, -\omega, \zeta$ the order types of respectively $(\mathbb{N}, <)$, $(-\mathbb{N}, <)$ and $(\mathbb{Z}, <)$. Below, we do not distinguish between a linear ordering and its order type unless necessary. This is safe since all the constructions we perform are defined up to similar order type.

A *subordering* β of α is a subset of L equipped with the same ordering relation; i.e., $\beta = (L', <)$ with $L' \subseteq L$. We write $\beta \subseteq \alpha$. A *convex subset of* α is a subset S of α such that for all $x, y \in S$ and $x < z < y$, $z \in S$. We use the notations $[x, y], [x, y[,]x, y],]x, y[,]-\infty, y],]-\infty, y[, [x, +\infty[$ and $]x, +\infty[$ for denoting the usual *intervals*. Intervals are convex, but the converse does not hold in general if α is not complete (see below). Given two subsets X, Y of a linear ordering, $X < Y$ holds if for all $x \in X$ and $y \in Y$, $x < y$.

The *sum* of two linear orderings $\alpha_1 = (L_1, <_1)$ and $\alpha_2 = (L_2, <_2)$ (up to renaming, assume L_1 and L_2 disjoint), denoted $\alpha_1 + \alpha_2$, is the linear ordering $(L_1 \cup L_2, <)$ with $<$ coinciding with $<_1$ on L_1, with $<_2$ on L_2 and such that $L_1 < L_2$. More generally, given a linear ordering $\alpha = (L, <)$ and for each $x \in L$ a linear ordering $\beta_x = (K_x, <_x)$ (the K_x are assumed disjoint), we denote by $\sum_{x\in\alpha} \beta_x$ the linear $(\cup_{x\in L} K_x, <')$ with $x' <' y'$ if $x < y$ or $x = y$ and $x' <_x y'$, where $x' \in K_x$ and $y' \in K_y$.

A linear ordering α is *complete* if every nonempty subset of α with an upper bound has a least upper bound in α, and every nonempty subset of α with a lower bound has a greatest lower bound in α.

A (Dedekind) *cut* of a linear ordering $\alpha = (L, <)$ is a couple (E, F) where $\{E, F\}$ is a partition of L, and $E < F$. Cuts are totally ordered by $(E, F) < (E', F')$ if $E \subsetneq E'$. This order has a minimal element $\bot = (\emptyset, L)$ and a maximal element $\top = (L, \emptyset)$. We denote by $\overline{\alpha}$ the set of cuts of α, and we abbreviate by $\overline{\alpha}^{]}, \overline{\alpha}^{[[}, \overline{\alpha}^{]]}, \overline{\alpha}^{][}$ the sets $\overline{\alpha}, \overline{\alpha}\setminus\{\top\}, \overline{\alpha}\setminus\{\bot\}, \overline{\alpha}\setminus\{\bot, \top\}$ respectively. An important remark is that $\overline{\alpha}$ is a complete linear ordering. Cuts can be thought as new elements located between the elements of α: given $x \in \alpha$, $x^- =]-\infty, x[, [x, +\infty[)$ represents the cut placed just before x, while $x^+ =]-\infty, x],]x, +\infty[)$ is the cut placed just after x. We say in this case that x^+ is *the successor of* x^- *through* x. But not all cuts are successors or predecessors of another cut. A cut c is a *right limit* (resp. a *left limit*) if it is not the minimal element and not of the form x^+ for some x in α (resp. not the maximal element and not of the form x^-).

A linear ordering α is *dense* if for every $x < y$ in α, there exists z in $]x, y[$. A linear ordering is *scattered* if none of its subordering is dense. For instance $(\mathbb{Q}, <)$ and $(\mathbb{R}, <)$ are dense, while $(\mathbb{N}, <)$ and $(\mathbb{Z}, <)$ are scattered. Being scattered is preserved under taking a subordering. A scattered sum of scattered linear orderings also yields a scattered linear ordering. Every ordinal is scattered. Furthermore, if α is scattered, then $\overline{\alpha}$ is scattered. And if α is countable and scattered, then $\overline{\alpha}$ is also countable and scattered.

Additional material on linear orderings can be found in [13].

2.2 Words, languages

We use a generalised version of words: words indexed by linear orderings. Given a linear ordering $\alpha = (L, <)$ and a finite alphabet A, an *α-word u over the alphabet A* is a mapping from L to A. We also say that α is the *domain* of the word u, or that u is a word *indexed* by α. Below we always consider word up to isomorphism of the domain, unless a specific presentation of the domain is required. Standard finite words are simply the words indexed by finite linear orderings. Given a word u of domain α and $\beta \subseteq \alpha$, we denote by $u|_\beta$ the word u restricted to its positions in β. Given an α-word u and a β-word v, uv represents the $(\alpha + \beta)$-word defined by $(uv)(x)$ is $u(x)$ if x belongs to α and $v(x)$ if x belongs to β. The product is extended to languages of words in a natural way. The product of words is naturally generalised to the infinite product $\prod_{i\in\alpha} u_i$, where α is an order type and u_i are linear β_i-words; the resulting being a $(\sum_{i\in\alpha} \beta_i)$-word.

For a language W and a linear ordering α, one defines W^α to be the language containing all the words $\prod_{i\in\alpha} u_i$, where $u_i \in W$ for all $i \in \alpha$.

Given an alphabet A, we denote by $A^\diamond$ the set of words indexed by a countable scattered linear ordering.

2.3 Semigroups and additive labellings

For a thorough introduction to semigroups, we refer the reader to [8, 9]. A *semigroup* $(S, .)$ is a set S equipped with an associative binary operator written multiplicatively. Groups and monoids are particular instances of semigroups. The set of nonempty finite words A^+ over an alphabet A is a semigroup – it is the semigroup freely generated by A. A *morphism of semigroups* from a semigroup $(S, .)$ to a semigroup $(S', .')$ is a mapping φ from S to S' such that for all x, y in S, $\varphi(x.y) = \varphi(x).'\varphi(y)$. An *idempotent* in a semigroup is an element e such that $e^2 = e$.

Let α be a linear ordering and $(S, .)$ be a semigroup. A mapping σ from couples (x, y) with $x, y \in \alpha$ and $x < y$ to S is called an *additive labelling* if for every $x < y < z$ in α, $\sigma(x, y).\sigma(y, z) = \sigma(x, z)$.

Given a semigroup morphism φ from $(A^\diamond, .)$ to some semigroup $(S, .)$ and a word u in $A^\diamond$ of domain α, there is a natural way to construct an additive labelling φ_u from $\overline{\alpha}$ to $(S, .)$: for every two cuts $x < y$ in $\overline{\alpha}$, set $\varphi_u(x, y)$ to be $\varphi(u_{x,y})$, where $u_{x,y}$ is the word u restricted to its positions between x and y; i.e., $u_{x,y} = u|_{F\cap E'}$ for $x = (E, F)$ and $y = (E', F')$.

3 Factorisation forest theorems

In this section, we present various theorems of factorisation forest. We first give the original statement in Section 3.1. In Section 3.2, we introduce the notion of a split, and use it in a different presentation of the result. In Section 3.3, we state the extension to every complete linear ordering.

3.1 Factorisation forest theorem

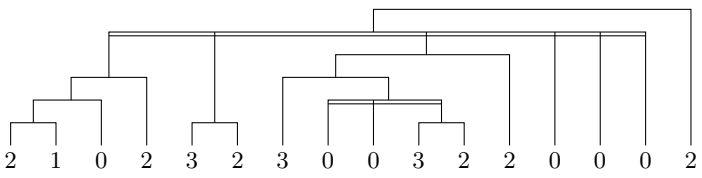

Fig. 1. A factorisation tree

Fix an alphabet A and a semigroup morphism φ from A^+ to a finite semigroup $(S, .)$. A *factorisation tree* is an ordered unranked tree in which each node is either a leaf labelled by a letter, or an internal node. The *value* of a node is the word obtained by reading the leaves below from left to right. A *factorisation tree* of a word $u \in A^+$ is a factorisation tree of value u. The *height* of the tree is defined as usual, with the convention that the height of a single leaf is 0. A factorisation tree is *Ramseyan* (for φ) if every node 1) is a leaf, or 2) has two children, or, 3) the values of its children are all mapped by φ to the same idempotent of S.

Example 1. Fix $A = \{0, 1, 2, 3, 4\}$, $(S, .) = (\mathbb{Z}/5\mathbb{Z}, +)$ and φ to be the only semigroup morphism from A^+ to $(S, .)$ mapping each letter to its value. Figure 1 presents a Ramseyan factorisation tree for the word $u = 210232300322002$. In this drawing, internal nodes appear as horizontal lines. Double lines correspond to case 3 in the description of Ramseyanity.

The theorem of factorisation forests is then the following.

Theorem 1 (factorisation forests, Simon [15]). *For every alphabet A, finite semigroup $(S, .)$, semigroup morphism φ from A^+ to S and word u in A^+, u has a Ramseyan factorisation tree of height at most $3|S|$.*

The original theorem is due to Simon [15], with a bound of $9|S|$. An improved bound of $7|S|$ is provided by Chalopin and Leung [6]. The value of $3|S|$ is a byproduct of the present work (see Theorem 2 below and subsequent comments).

3.2 A variant via Ramseyan splits

The variant presented here of the factorisation forest theorem uses the notion of splits. We reuse this framework later on.

A *split of height N* of a linear ordering α is a mapping s from α to $[1, N]$. Given a split, two elements x and y in α such that $s(x) = s(y) = k$ are *k-neighbours* if $s(z) \geq k$ for all $z \in [x, y]$. k-neighbourhood is an equivalence relation over $s^{-1}(k)$. Fix an *additive labelling* from α to some finite semigroup S. A split of α is *Ramseyan* for σ — we also say a *Ramseyan split for* (α, σ) — if for every $k \in [1, N]$, every $x < y$ and $x' < y'$ such that all the elements x, y, x', y' are k-neighbours, then $\sigma(x, y) = \sigma(x', y') = (\sigma(x, y))^2$; Equivalently, for all k, every class of k-neighbourhood is mapped by σ to a single idempotent.

Example 2. Let S be $\mathbb{Z}/5\mathbb{Z}$ equipped with the addition $+$. Consider the linear ordering of 17 elements and the additive labelling σ defined by:

$$| 3 | 1 | 0 | 2 | 3 | 2 | 3 | 0 | 0 | 3 | 2 | 2 | 0 | 0 | 0 | 2 |$$

Each symbol '|' represents an element, the elements being ordered from left to right. Between two consecutive elements x and y is represented the value of $\sigma(x, y) \in S$. In this situation, the value of $\sigma(x, y)$ for every $x < y$ is uniquely

defined according to the additivity of σ: it is obtained by summing all the values between x and y modulo 5.

A split s of height 3 is the following, where we have written above each element x the value of $s(x)$:

$$\begin{array}{cccccccccccccccccc} 1 & & 3 & & 2 & & 2 & & 1 & & 2 & & 1 & & 2 & & 2 & & 2 \\ | & 2 & | & 1 & | & 0 & | & 2 & | & 3 & | & 2 & | & 3 & | & 0 & | & 0 & | & 3 & | \end{array}$$

$$\begin{array}{cccccccccccccccc} 3 & & 2 & & 2 & & 1 & & 1 & & 1 & & 1 & & 2 \\ 2 & | & 2 & | & 0 & | & 0 & | & 0 & | & 2 & | \end{array}$$

In particular, if you choose $x < y$ such that $s(x) = s(y) = 1$, then the sum of elements between them is 0 modulo 5. If you choose $x < y$ such that $s(x) = s(y) = 2$ but there is no element z in between with $s(z) = 1$ — i.e., x and y are 2-neighbours — the sum of values separating them is also 0 modulo 5. Finally, it is impossible to find two distinct 3-neighbours in our example.

Theorem 2. *For every finite linear ordering α, every finite semigroup $(S, .)$ and additive labelling σ from α to S, there exists a Ramseyan split for (α, σ) of height at most $|S|$.*

Let us state the link between Ramseyan splits and factorisation trees. Fix an alphabet A, a semigroup S, a morphism φ from A^+ to S and a word $u \in A^+$ of finite domain α. The following is easy to establish:

- every Ramseyan factorisation tree of height k of u can be turned into a Ramseyan split of height at most k of $(\overline{\alpha}^{]\![}, \varphi_u)$,
- every Ramseyan split of height k of $(\overline{\alpha}^{]\![}, \varphi_u)$ can be turned into a factorisation tree of height at most $3k$ of u.

Using this last argument and Theorem 2, we directly obtain a proof of Theorem 1 with the announced bound of $3|S|$.

3.3 Ramseyan splits for complete linear orderings

We generalise Theorem 2 to complete linear orderings as follows.

Theorem 3. *For every complete linear ordering α, every finite semigroup $(S, .)$ and additive labelling σ from α to S, there exists a Ramseyan split for (α, σ) of height at most $3|S|$ ($|S|$ if α is an ordinal).*

Compared to Theorem 2, we trade the finiteness — which is replaced by the completeness — for a bound of $3|S|$ — which replaces a bound of $|S|$. The special case of α being a finite ordinal yields Theorem 2.

The proof by itself follows the lines of [6]. This means using three different arguments according to three different situations arising in the decomposition of the semigroup by Green's relations. The first situation amounts to treat the case of S being a group. The second case is the one of a single $\mathcal{J}$-class ($\mathcal{J}$ is one of the Green's relation). And finally one performs an induction on the number of $\mathcal{J}$-classes. Examples 1 and 2 do only involve the first situation.

This rough sketch contains certain technicalities when the proof need be formalised. In particular one performs many gluing and nesting of splits. An

explanation of the improvement of the bound in the finite case is that splits are more versatile in handling those details. E.g., the use of the 'border types' $[[,][,]], []$ allow to glue more easily pieces of Ramseyan splits together, while Ramseyan factorisation trees do correspond only to the case $][$.

4 Application to countable scattered linear orderings

In this section, we present the automata theoretic approach to regularity of languages over words of countable scattered domain. This notion has been developed in [3], in which a suitable family of automata is proposed. These automata are easily shown closed under union, intersection and projection, and their emptiness is decidable. The closure under complementation is more involved and is due to Carton and Rispal [5]. In this section we give a simplified proof to this result.

The properties of these automata result directly in the decidability of the monadic (second-order) theory of countable scattered linear orderings. This decidability result can be independently established using the famous theorem of Rabin [12] (see [16] for a modern presentation), and its consequence, the decidability of the monadic theory of $(\mathbb{Q}, <)$. But this technique is less informative and is not totally satisfying. More precisely, using the theorem of Rabin signifies the use of infinite trees, and also has to do with the theory of Müller/parity games and their determinacy. We believe that these subtle issues are not relevant when considering the theory of linear orderings, and thus are worth being avoided. Furthermore the approach using the theorem of Rabin does not help much for understanding the notions of regularity over linear orderings.

Another application of Theorem 3 – to some extent a variant of the application proposed here – is to give a compositional proof for the decidability of the monadic theory of countable scattered linear orderings. Generally speaking, the compositional method allows to devise automata-free proofs of decidability of monadic theories (or other logics). It was used by Shelah [14] in his seminal work on the monadic theory of linear orderings. But so far it could not be used in situations like scattered orderings by lack of the correct combinatorial result. Theorem 3 bridges this gap.

In this section we concentrate ourselves solely on the technical core of the theory: the closure under complementation of automata over countable scattered linear orderings. We present the suitable family of automata, then the corresponding semigroup, and finally the complementation proof itself.

4.1 Automata over countable scattered linear orderings

In this section, we define priority automata and show how they accept words indexed by countable scattered linear orderings. Those automaton were introduced in [3], but in their 'Muller' form, while here we adopt the 'parity-like' approach (to this respect, the results given below are new).

Definition 1. *A* priority automaton $\mathcal{A} = (Q, A, I, F, p, \delta)$ *consists of a finite set of* states Q*, a finite alphabet* A*, a set of* initial states I*, a set of final states* F*,*

a priority mapping $p : Q \mapsto [1, N]$ *(N being a natural) and a* transition relation $\delta \subseteq (Q \times A \times Q) \uplus ([1, N] \times Q) \uplus (Q \times [1, N])$.

A *run* of the automaton $\mathcal{A}$ over an α-word u is a mapping ρ from $\overline{\alpha}$ to Q such that for all cuts c, c':

- if c' is the successor of c through x, then $(\rho(c), u(x), \rho(c')) \in \delta$,
- if c is a right limit, then $(k, \rho(c)) \in \delta$ where $k = \max \bigcap_{c'<c} p(\rho(]c', c[))$,
- if c is a left limit, then $(\rho(c), k) \in \delta$ where $k = \max \bigcap_{c'>c} p(\rho(]c, c'[))$.

The first case corresponds to standard automata on finite words: a transition links one state to another while reading a single letter in the word. The second case verifies that the highest priority appearing infinitely close to the left of c corresponds to a transition. The third case is symmetric. An α-word u is *accepted* by $\mathcal{A}$ if there is a run ρ of $\mathcal{A}$ over u such that $\rho(\bot) \in I$ and $\rho(\top) \in F$.

Example 3. Consider the automaton with states $\{q, r\}$, alphabet $\{a\}$, initial states $\{q, r\}$, final state q, priority mapping constant equal to 0 and transitions $\{(q, a, q), (q, a, r), (0, q), (r, 0)\}$). It accepts those words in $\{a\}^\diamond$ which have a complete domain. For this, note that a linear ordering is complete iff no cut is simultaneously a left and a right limit.

Consider a word $u \in \{a\}^\diamond$ which has a complete domain α. For $c \in \overline{\alpha}$, set $\rho(c)$ to be q if c is $\top$ or if c has a successor, else $\rho(c)$ is r. Under the hypothesis of completeness, it is simple to verify that ρ is a run witnessing the acceptance of the word. Conversely, assume that there is a run ρ over the α-word u with α not complete. There is a cut $c \in \overline{\alpha}$ which is both a left and a right limit. If $\rho(c)$ is r, then, as c is a left limit, there is no corresponding transition; else if $\rho(c)$ is q the same argument can be applied to the right of c. In both cases there is a contradiction.

It is easy to prove that the languages of $\diamond$-words accepted by priority automata are closed under union, intersection, and projection [5]. It is also easy to establish the decidability of their emptiness problem. Below, after introducing the necessary semigroup, we show the more difficult closure under complementation.

4.2 Semigroup structure

In order to use Theorem 3, we have to relate automata with semigroups. Let us fix ourselves an automaton of states Q and priorities $[1, N]$. One equips

$$S = 2^{Q \times [1,N] \times Q}$$

of a semigroup structure as usual with

$$\text{for } a, b \in S, \qquad a.b = \{(p, \max\{m, n\}, r) \ : \ (p, m, q) \in a,\ (q, n, r) \in b\} \ .$$

This definitions naturally comes together with a semigroup morphism φ from $\diamond$-words to S such that for every word u, $\varphi(u)$ contains (p,n,q) iff there exists a run of the automaton reading u, starting from state p, finishing with state q, and of maximal priority n.

The semigroup defined so far does not entirely capture the semantic of the automaton. In particular it contains no limit passing features. We resolve this issue by defining the exponentiations under ω and $-\omega$ of idempotents of the semigroup. One defines e^ω (and symmetrically $e^{-\omega}$) for an idempotent e by:

$$e^\omega = e.\{(q,m,r) \ : \ (q,m,q) \in e,\, (\max(m,p(q)),r) \in \delta\}\ ,$$
$$\text{and} \quad e^{-\omega} = \{(r,m,q) \ : \ (q,m,q) \in e,\, (r,\max(m,p(q))) \in \delta\}.e\ .$$

One also defines e^ζ as $e^{-\omega}.e^\omega$.

The essential property of these exponentiations is the following. Given a sequence of words $(u_i)_\beta$ indexed by $\beta = \omega, -\omega, \zeta$, and such that for all i in β, $\varphi(u_i) = e$, then

$$\varphi(\prod_{i\in\beta} u_i) = e^\beta\ .$$

4.3 Complementation

We sketch now a short proof of the following theorem.

Theorem 4 (Carton and Rispal [5]). *Languages of countable scattered words accepted by priority automata are closed under complement.*

Let k be a natural number, a be in S, and ι among $[], [[,]],][$, set $S_k^\iota(a)$ to be the set of $\diamond$-words u such that $\varphi(u) = a$ and $(\overline{\alpha}^\iota, \varphi_u)$ admits a Ramseyan split of height k (by convention, ε does not belong to $S_k^{][}(a)$). We prove by induction on k that for every a in S and $\iota = [], [[,]],][$, $S_k^\iota(a)$ is accepted by a priority automaton. Since by Theorem 3, $\varphi^{-1}(a) = S^{[]}_{3|S|}(a)$, we deduce that $\varphi^{-1}(a)$ would be accepted by a priority automaton. As the complement language we are aiming at is a finite union of such languages, it would also be accepted by a priority automaton. This argument concludes the proof. What remains to be done is to establish the induction.

The base case is obtained by remarking that the following languages are accepted by priority automata:

$$S_0^{][}(a) = \varphi^{-1}(a) \cap A\,, \quad S_0^{[[}(a) = S_0^{]]}(a) = \varphi^{-1}(a) \cap \{\varepsilon\}\,, \quad \text{and} \quad S_0^{[]}(a) = \emptyset\,.$$

For all $k \geq 1$ and idempotent e, let $C_{e,k}$ be the set of $\diamond$-words u of domain α such that $\varphi(u) = e$, and there exists a split s of height k of $\overline{\alpha}$ such that $s(\bot) = s(\top) = 1$.

Our first step is to show how to construct an automaton accepting $C_{e,k+1}$ from automata accepting the languages $S_k^\iota(a)$. For this, consider the following

languages:

$$M_{e,k} = S_k^{][}(e)\,, \qquad M_{e,k}^{\leftarrow} = \sum_{ae^{-\omega}=e} S_k^{]]}(a),$$

$$M_{e,k}^{\rightarrow} = \sum_{e^{\omega}a=e} S_k^{[[}(a), \qquad \text{and } M_{e,k}^{\rightarrow\leftarrow} = \sum_{e^{\omega}ae^{-\omega}=e} S_k^{[]}(a).$$

By induction hypothesis, those languages are accepted by priority automata. Wlog, we choose them to use distinct priorities, and we set $n-1$ to be the maximal priority involved in those automata. We use them in the construction of the automaton $\mathcal{A}_{e,k+1}$ depicted Fig. 2.

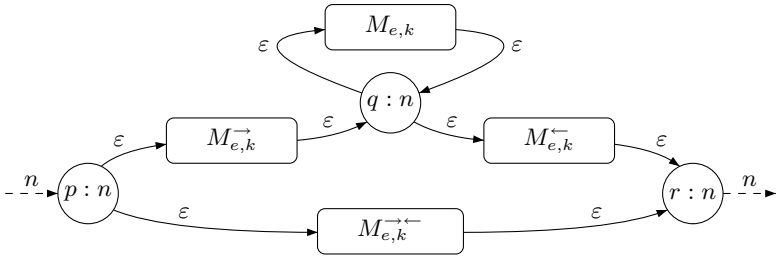

Fig. 2. The automaton $\mathcal{A}_{e,k+1}$

This construction makes use of ε-transitions. This is just a commodity of notation and can be removed using standard techniques. The automaton itself is made of disjoint copies of the automata accepting $M_{e,k}, M_{e,k}^{\rightarrow\leftarrow}, M_{e,k}^{\rightarrow}$, and $M_{e,k}^{\leftarrow}$, together with three new states p, q, r. Each ε-transition entering one of the subautomata represents in fact all possible ε-transitions with an initial state as destination; similarly, every ε-transition exiting a subautomaton represents all possible ε-transitions with as origin any of the final states of the automaton. The priority of the new state q is n, a priority unused elsewhere by construction. One chooses also p and r to have priority n (this is not of real importance since it is impossible to see infinitely often p or r in a run without seeing infinitely often q: the priority of q only matters). The two dashed arrows represent the two limit transitions (n, p) and (r, n).

Let $L_{e,k+1}[q_1, q_2]$ be the language accepted by this automaton with initial state q_1 and final state q_2 for q_1, q_2 among p, q, r.

The core of the proof is embedded in the following lemma.

Lemma 1. *For every idempotent e, $L_{e,k+1}[q, q] = C_{e,k+1}$.*

Proof. (sketch of the difficult inclusion: $L_{e,k+1}[q, q] \subseteq C_{e,k+1}$)

Let u be in $L_{e,k+1}[q, q]$, we have to construct a Ramseyan split s of height $k+1$ of $\varphi_u^{[]}$ with $s(\bot) = s(\top) = 1$. Since $u \in L_{e,k}[q, q]$, there exists a corresponding

run ρ of the automaton $\mathcal{A}_{e,k+1}$ from state q to state q. Let I be the set of cuts c such that $\rho(c) = q$.

Set $s(c) = 1$ for all c in I. Let now $J \subseteq \overline{\alpha}$ be a maximal interval not intersecting I. Let us define s over J. Let x be $\inf J$ and y be $\sup J$, J is either $[x, y]$, $[x, y[$, $]x, y]$ or $]x, y[$. Assume $J = [x, y[$. In this case, since $y \notin J$, $y \in I$ and hence $\rho(y) = q$. Furthermore, since $x \in J$, there exists an infinite sequence $x_1 < x_2 < \ldots$ of length ω and limit x in I. As the priority of $\rho(x_i) = q$ is the maximal one, namely n, the only possible state for $\rho(x)$ compatible with limit transitions is p. Furthermore the state q is never visited by ρ in $[x, y[$ (by definition of J). By inspecting the automaton, we conclude that the only possibility is that ρ restricted to $[x, y]$ is in fact a run of the subautomaton $M^{\rightarrow}_{e,k}$. By induction hypothesis, since $M^{\rightarrow}_{e,k}$ is a union of languages $S^{[[}_k$, this means that there exists a split s_J of height k of J, Ramseyan for σ. We set s to coincide with $s_J + 1$ over J. For the other possibilities for J, runs of the automata $M^{\rightarrow}_{e,k}$, $M^{\leftarrow}_{e,k}$ and $M^{\rightarrow\leftarrow}_{e,k}$ are involved in a similar way.

Proving the correctness of this construction requires some more arguments. Let us come back to the case $J = [x, y[$ above. The run ρ over $[x, y]$ together with the definition of $M^{\rightarrow}_{e,k}$ witnesses that $e^{\omega}\sigma(x, y) = e$. This is a *local correctness property* for the construction. What we have to prove is that $\sigma(x, y) = e$ for every $x < y$ in I; i.e., a *global correctness conclusion*. This propagation of the local equalities to the global level is achieved using topological arguments. In particular, it uses the scatteredness hypothesis over α as well as the countability hypothesis. It also involves the use the countable axiom of choice. □

We can derive from the last lemma the following.

Corollary 1. *$L_{e,k}[q, p] = C^{\omega}_{e,k}$, $L_{e,k}[r, q] = C^{-\omega}_{e,k}$, and $L_{e,k}[r, p] = C^{\zeta}_{e,k}$.*

And we terminate by remarking that, for $\iota = [], [[,]],][$ and $a \in S$, the language $S^{\iota}_k(a)$ can be written in terms of the S_{k-1} and the $C_{e,k}$ languages using finite sums, concatenation and ω, $-\omega$ and ζ exponentiations.

Conclusion and future work

We believe that the factorisation forest theorem cannot be improved further in the directions presented here. In particular, the bounds in Theorem 2 cannot be improved in general. And in Theorem 3, removing the completeness hypothesis makes the result fail.

Concerning automata over countable scattered linear orderings, our complementation proof has the advantage – with respect to the original one in [5] – to isolate the combinatorial part from the problems related to scatteredness itself. Our proof is in fact very resemblant to the original one of Buchi for ω-words [4] in which the theorem of Ramsey would be replaced by Theorem 3. Along the same lines, Theorem 3 can also be used in a compositional proof of the decidability of the monadic theory of countable scattered linear orderings.

The question is whether there are other applications for Theorem 3 since $(\mathbb{R}, <)$ does not have a decidable monadic theory [14]. We believe that it is the case, for instance for tackling the conjecture of Rabin that the monadic theory of $(\mathbb{R}, <)$ is decidable when monadic variables are interpreted over Borelian sets (let us remark that the theory of $(\mathbb{R}, <)$ with quantification over boolean combinations of opens sets is already known to be decidable from Rabin [12]). We are working in this direction.

Acknowledgement I am very grateful to Olivier Carton for his numerous comments on this work. I also thank the anonymous referees who helped in improving this document.

References

1. M. Bojańczyk and T. Colcombet. Bounds in omega-regularity. In *IEEE Symposium on Logic In Computer Science*, pages 285–296, 2006.
2. T. C. Brown. An interesting combinatorial method in the theory of locally finite semigroups. *Pacific Journal of Mathematics*, 36(2):277–294, 1971.
3. V. Bruyère and O. Carton. Automata on linear orderings. In *MFCS*, volume 2136, pages 236–247, 2001.
4. J. R. Büchi. On a decision method in restricted second order arithmetic. In *Proceedings of the International Congress on Logic, Methodology and Philosophy of Science*, pages 1–11. Stanford University press, 1960.
5. O. Carton and C. Rispal. Complementation of rational sets on countable scattered linear orderings. *Int. J. Found. Comput. Sci.*, 16(4):767–786, 2005.
6. J. Chalopin and H. Leung. On factorization forests of finite height. *Theoretical Computer Science*, 310(1–3):489–499, jan 2004.
7. T. Colcombet. A combinatorial theorem for trees. In *ICALP'07*, Lecture Notes in Computer Science. Springer-Verlag, 2007.
8. G. Lallement. *Semigroups and Combinatorial Applications*. Wiley, New-York, 1979.
9. J. Pin. *Varieties of formal languages*. North Oxford, London and Plenum, New-York, 1986.
10. J. Pin, B. le Saëc, and P. Weil. Semigroups with idempotent stabilizers and application to automata theory. *Int. J. of Alg. and Comput.*, 1(3):291—314, 1991.
11. J.-E. Pin and P. Weil. Polynominal closure and unambiguous product. *Theory Comput. Syst.*, 30(4):383–422, 1997.
12. M. Rabin. Decidability of second-order theories and automata on infinite trees. *Trans. Amer. Math. soc.*, 141:1–35, 1969.
13. J. G. Rosenstein. *Linear Orderings*. Academic Press, New York, 1982.
14. S. Shelah. The monadic theory of order. *Annals Math*, 102:379–419, 1975.
15. I. Simon. Factorization forests of finite height. *Theor. Comput. Sci.*, 72(1):65–94, 1990.
16. W. Thomas. Languages, automata, and logic. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Language Theory*, volume III, pages 389–455. Springer, 1997.